

Prior Distributions for Ranking Problems

Toby Kenney

Hao He

Hong Gu

October 28, 2016

Abstract

The ranking problem is to order a collection of units by some unobserved parameter, based on observations from the associated distribution. This problem arises naturally in a number of contexts, such as business, where we may want to rank potential projects by profitability; or science, where we may want to rank variables potentially associated with some trait by the strength of the association. Most approaches to this problem are empirical Bayesian, where we use the data to estimate the hyperparameters of the prior distribution, then use that distribution to estimate the unobserved parameter values. There are a number of different approaches to this problem, based on different loss functions for mis-ranking units. However, little has been done on the choice of prior distribution. Typical approaches involve choosing a conjugate prior for convenience, and estimating the hyperparameters by MLE from the whole dataset. In this paper, we look in more detail at the effect of choice of prior distribution on Bayesian ranking. We focus on the use of posterior mean for ranking, but many of our conclusions should apply to other ranking criteria, and it is not too difficult to adapt our methods to other choices of prior distributions.

keywords

Empirical Bayes; posterior mean ranking; choice of prior

1 Introduction

Suppose we have a collection of units we want to rank by a certain feature of each unit: for example, we may wish to rank genes by the risk they cause of a particular condition; we may wish to rank sportsmen by their success-rate at particular standardised trials; we may wish to rank business opportunities by the profit they will generate. This is a very common inference problem first studied as a formal statistical problem by Bechhofer (1954) and by Gupta (1956). Typically, for each unit we wish to rank, we will have some data on the associated feature, but will not know the true value of that feature. Based on our data, we will have a point estimate for the feature, and an associated error distribution. The amount of data we might have for different units can vary wildly, meaning that the associated error distributions can be very different for different units. This means that when we select the top units using only our point estimates, the units for which we have largest errors have a higher chance of appearing among the top units, because a large error increases the chance of the point estimate being large. We are therefore likely to select a large number of false positives if we select based solely on the point estimates.

We can illustrate this with a simple example. Suppose we have 300 coins, we toss 100 of them six times each, 100 of them eight times each, and the remaining 100 of them ten times each, and rank the coins by the proportion of heads observed. If the coins are all fair, then among the 100 that we toss six times each, there is likely to be at least one that achieves 100% heads. Among the 100 that we toss eight times, there might be one that achieves 100% heads, and there are likely to be several that achieve 87.5% heads. Among the 100 that we toss ten times each, it is fairly

unlikely than any will exceed 80%, so the highest ranked units will almost certainly come from among the coins that we toss only six times. That is, the highest ranked units are almost all false-positives arising only out of chance. This is still true, even if one or more of the coins that are tossed ten times have a slightly higher probability of heads.

On the other hand, if our main aim is to avoid false positives, we could use a testing-based approach, where for each unit, we perform an hypothesis test of whether the unit has some null status — for example whether the probability of heads is 0.5. We can then rank by the p -values of these tests. This has the advantage of minimising false positives, but in many cases there are a large number of true positives, but only a few of them are truly important. If we apply the testing approach, we will often select the units on which we have collected most data, simply because the more data we have, the more evidence that they are not null cases. This may lead to neglecting some units which have much higher underlying value, but for which we have less data.

Other approaches to the problem mainly take a Bayesian approach. They assume that the true values of the relevant feature fall under some distribution. We can estimate this underlying distribution from all the data points. Then for each observed unit, we use this distribution as a prior to estimate a posterior distribution of the true value for this unit. We then perform our ranking based on these posterior distributions and a choice of loss function. There are a range of different methods based on different loss functions. For example, posterior expected rank (Laird and Louis, 1989) use a loss function the squared difference between the true rank of a unit (based on the actual value of the feature) and the estimated rank. The r -values method (Henderson and Newton, 2015), corresponds to a loss function the sum of absolute differences between estimated rank and true rank. Both of these loss functions are based entirely upon ranks, with no consideration of the actual true values. That is, they consider mis-ranking two units with almost identical true values to be as bad as mis-ranking units with very different true values. For the vast majority of practical ranking problems, this will not be the case. Gelman and Price (1999) present the interesting case of looking for spatial patterns among the top-ranked units, where artificial patterns can arise from patterns in available sample sizes. For their purposes, the ideal ranking method would be in such a way that the distribution of rank is the same for all values of standard error. For a known prior, it is possible to calculate this rank, though we are not aware of any work applying such a ranking method. However, methods with loss functions based only on rank, rather than value might be expected to perform better on this criterion, since all errors in ranking can cause this issue equally.

The aim of a ranking analysis is often to maximise the average true value from the selected units. For instance, in the business profit example, the aim would be to maximise the expected total profit. For these purposes, the loss function is the difference between the largest true values and the true values of selected units. This loss function is introduced in Gupta and Hsiao (1983), with some additional thought given to the situation where the loss is different for the case of omitting a variable that should be included, from the case of including a variable that should be omitted. They show that for this loss function with known prior the Bayes rule is to rank by posterior mean (though they are not very explicit about this, and include some unnecessary hypotheses). This posterior mean ranking is used for example in Aitkin and Longford (1986). A range of other loss functions have also been considered, for example, Lin *et al.* (2006) summarise a range of choices of loss function. For this paper, we will be focussing on the posterior mean ranking method, and its corresponding loss function, although many of our methods can be easily adapted to other Bayesian ranking methods.

The key difficulty in Bayesian ranking methods is to choose the form of the prior. Two common choices are the conjugate prior (which for normal error is normal), and a non-parametric prior, which can be calculated using the results of Laird (1978). Figure 1 shows the sort of problem that can arise with this approach. The lines on that figure show points that are ranked equally by posterior mean under a normal prior estimated from the whole dataset. As can be seen in that plot, a lot of emphasis gets placed on points with small variance. The reason is that the normal prior is light-tailed, so large true values are deemed implausible, and discounted. However, the true prior distribution seems to be more heavy-tailed than the normal, so larger values should not be discounted so much. For example, consider the point in the red circle. While it does have a larger standard error, it is very significantly non-zero, and it is likely that the true log-odds ratio is high. Intuitively, we would probably want to rank this data point among the

very top-ranked units. However, the posterior mean under the normal prior ranks it below a lot of other points which, while certainly significantly non-zero, have very small effect size. For practical purposes, this is not desirable. We are usually interested in units with a large effect size.

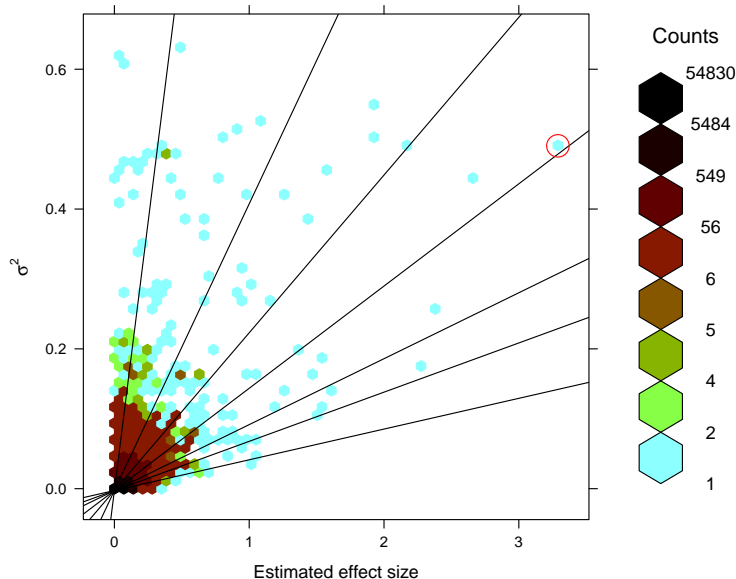


Figure 1: Estimated log-odds ratio versus variance of estimator for SNP data from a GWAS study into type-2 diabetes (Morris *et al.*, 2012). The lines show points ranked equally under posterior mean with a normal prior.

The aim of this paper is to study the effect that choice of prior can have on the ranking problem, and determine suitable choices of prior for such analyses. Despite a fair amount of literature on Bayesian ranking methods, there has been a noticeable lack of work on the question of choice of prior. In view of the fact that selecting a suitable model for the prior distribution is a very difficult problem in model selection, it is important to consider the effects of a misspecified prior distribution. As will become apparent later, certain choices of prior are inherently more robust to misspecification than others. Furthermore, some choices of prior are more sensitive to parameter estimation than others.

We describe the objective more formally as follows. A ranking problem consists of a collection of units with unobserved parameters θ_i . For each unit, we have a point estimate x_i for θ_i . We assume that x_i is normally distributed with mean θ_i and variance σ_i^2 , where σ_i is known. It is straightforward to adapt our approach to a number of other error distributions, but for this paper, we will focus on the normal error case. We assume that the unobserved values θ_i follow what we will refer to as the *true prior* distribution. We will rank by posterior mean using what we will refer to as the *estimating prior*, which may or may not be the same as the true prior. We are interested in how choice of the estimating prior affects the ranking.

The structure of this paper is as follows: In Section 2, we develop some theory behind posterior mean ranking, and the loss from using the wrong estimating prior. In Section 3, we give a visual representation of the effect of choice of estimating prior on posterior mean ranking. In Section 4, we show that using the non-parametric MLE as an estimating prior for posterior mean ranking produces a robust ranking. In Section 5, we apply our theory to some

examples of misspecified estimating priors, and perform a simulation study to confirm the results are as expected. We show that an exponential estimating prior is a good general-purpose choice for posterior mean ranking. In Section 6, we apply this to some real data examples where we show the difference in the ranking between using a normal distribution for the estimating prior and using an exponential distribution. In Section 7, we make some concluding remarks and suggestions for further investigations.

2 Theory

2.1 Approximate Posterior Mean for given Prior Distribution

We suppose that our true prior distribution is continuous and has density function $\pi(\theta)$. Suppose that we have a point estimate x , whose error distribution is normal with variance σ^2 , where σ is small. Since σ is small, values of θ that are far from x are extremely implausible, and contribute little to the posterior mean for most choices of $\pi(\theta)$. We therefore focus on the form of $\pi(\theta)$ for values of θ close to x . Taking a first order Taylor expansion about x gives

$$\pi(\theta) = \pi(x) + \pi'(x)(\theta - x)$$

Using this approximation to $\pi(\theta)$ gives that the posterior mean is

$$\begin{aligned} \frac{\int (x + (\theta - x)) (\pi(x) + \pi'(x)(\theta - x)) e^{-\frac{(\theta-x)^2}{2\sigma^2}} d\theta}{\int (\pi(x) + \pi'(x)(\theta - x)) e^{-\frac{(\theta-x)^2}{2\sigma^2}} d\theta} &= x + \frac{\int (\theta - x) (\pi(x) + \pi'(x)(\theta - x)) e^{-\frac{(\theta-x)^2}{2\sigma^2}} d\theta}{\int (\pi(x) + \pi'(x)(\theta - x)) e^{-\frac{(\theta-x)^2}{2\sigma^2}} d\theta} \\ &= x + \frac{\pi(x) \int (\theta - x) e^{-\frac{(\theta-x)^2}{2\sigma^2}} d\theta + \pi'(x) \int (\theta - x)^2 e^{-\frac{(\theta-x)^2}{2\sigma^2}} d\theta}{\pi(x) \int e^{-\frac{(\theta-x)^2}{2\sigma^2}} d\theta + \pi'(x) \int (\theta - x) e^{-\frac{(\theta-x)^2}{2\sigma^2}} d\theta} \\ &= x + \frac{\pi'(x)}{\pi(x)} \sigma^2 \end{aligned}$$

This means that the key part of choice of estimating prior is to estimate the quantity $\lambda(x) = -\frac{\pi'(x)}{\pi(x)}$. For the tail of the distribution, this quantity is positive, and asymptotically approaches the hazard rate. For an exponential distribution, it is constant. For heavier-tailed distributions it tends to zero as $x \rightarrow \infty$. For light-tailed distributions, it tends to infinity as $x \rightarrow \infty$.

2.2 Loss function in terms of posterior misestimation

Suppose we should estimate the posterior mean as $x - \lambda\sigma^2$, but in fact, we estimate it as $x - \hat{\lambda}\sigma^2$, for some particular value of x . The question is what is the average loss function resulting from this. For a ranking of all the observations, we can consider the total loss as the sum of losses due to individual mis-rankings. That is, suppose we rank the observations $\theta_{[1]}, \theta_{[2]}, \dots, \theta_{[n]}$, when the correct ranking is $\theta_{(1)}, \theta_{(2)}, \dots, \theta_{(n)}$. If we choose our selection cutoff as the

first k units, then the loss function is

$$\begin{aligned}
I_k &= \sum_{i=1}^k (\theta_{(i)} - \theta_{[i]}) \\
&= \left(\sum_{\substack{i \leq k \\ \theta_{(i)} \notin \{\theta_{[1]}, \dots, \theta_{[k]}\}}} \theta_{(i)} \right) - \left(\sum_{\substack{j \leq k \\ \theta_{[j]} \notin \{\theta_{(1)}, \dots, \theta_{(k)}\}}} \theta_{[j]} \right)
\end{aligned}$$

We can move from the correct ranking to the estimated ranking by a series of transpositions of adjacent units in the current ranking. For example, if the correct ranking is 1, 2, 3, 4, 5, 6 and the estimated ranking is 2, 3, 1, 6, 5, 4, we can change from the correct ranking to the estimated ranking via the following sequence:

1 2 3 4 5 6
2 1 3 4 5 6
2 3 1 4 5 6
2 3 1 4 6 5
2 3 1 6 4 5
2 3 1 6 5 4

For each such transposition, exchanging the position of $\theta_{(i)}$ in the m th position, with $\theta_{(j)}$ in the $(m + 1)$ th position, the change in loss function is

$$\begin{cases} \theta_{(i)} - \theta_{(j)} & \text{if } m = k \\ 0 & \text{otherwise} \end{cases}$$

The total loss from this mis-ranking is then given by the sum of the loss functions for each transposition. We see that the loss for each transposition is non-negative for each value of k , so we can analyse the overall loss of a misranking by looking at the loss of each pairwise misranking.

If we consider the overall loss as the total of the loss functions for all values of k , we see that this loss function is just the sum of the loss functions for each transposition. Furthermore, whatever sequence of transpositions is performed, there will be one transposition for each misranked pair. Therefore the total loss function is the sum of the losses from each misranked pair. We can therefore study the total loss function by studying the misranking loss for any pair of observations. In practice, we will often consider only the loss of the upper tail of the distribution. That is, we will choose some cutoff a and evaluate the sum of the loss function for all k such that $x_{(k)} > a$. For this we have the following proposition (proof in Appendix A)

Proposition 2.1. *Suppose the true prior distribution of the parameter θ has density function $\pi(\theta)$, and that we have two observations x_1 and x_2 which are normally distributed with means θ_1 and θ_2 and standard deviations σ_1 and σ_2 respectively, where θ_1 and θ_2 are random samples from the true prior distribution, and σ_1 and σ_2 are assumed to be small.*

- (i) *The expected loss when the estimating prior and the true prior are the same (which we will refer to as the optimal expected loss) is approximately given by*

$$\frac{\sigma_1^2 + \sigma_2^2}{2} \mathbb{E}(\pi(x))$$

(ii) When the estimating prior has density $\hat{\pi}$, the difference between the expected loss and the optimal expected loss is approximately given by

$$\frac{1}{2}(\sigma_1^2 - \sigma_2^2)^2 \int_a^\infty \pi(x)^2 (\lambda(x) - \hat{\lambda}(x))^2 dx \quad (1)$$

where $\hat{\lambda}(x) = -\frac{\hat{\pi}'(x)}{\hat{\pi}(x)}$.

(iii) The difference between the expected loss from using the point estimate x_i and the optimal expected loss is approximately given by

$$\frac{1}{2}(\sigma_1^2 - \sigma_2^2)^2 \int_a^\infty \pi'(x)^2 dx$$

We see that for $\hat{\lambda}(x) \leq \lambda(x)$, $\int_a^\infty \pi(x)^2 (\lambda(x) - \hat{\lambda}(x))^2 dx$ is bounded by $\int_{-\infty}^\infty \pi(x)^2 \lambda(x)^2 dx$, which is the expected information of θ , and is bounded for most distributions. This means that if the estimating prior is too heavy-tailed, we can do no worse than ranking by point estimators alone. On the other hand, if we have $\hat{\lambda}(x) \geq \lambda(x)$, then the integral can approach $\int_{-\infty}^\infty \pi(x)^2 \hat{\lambda}(x)^2 dx$, which can be unbounded if the true prior has a heavy tail, but the estimating prior has a light tail. In most cases, the expression will not be unbounded. For example, for a normal estimating prior and a Pareto true prior, we have that $\hat{\lambda}(x) = \frac{x}{\tau^2}$ and $\pi(x) = \frac{\alpha \eta^\alpha}{x^{\alpha+1}}$, so

$$\int_0^\infty \pi(x)^2 \hat{\lambda}(x)^2 dx = \frac{\alpha^2 \eta^{2\alpha}}{\tau^4} \int_0^\infty x^{2-2(\alpha+1)} dx = \frac{\alpha^2 \eta^{2\alpha}}{\tau^4} \int_0^\infty x^{-2\alpha} dx$$

which diverges whenever $\alpha \leq \frac{1}{2}$. Thus for very heavy-tailed true priors, the loss from using a light-tailed estimating prior can diverge.

We see that there is a risk of this unbounded loss whenever $\hat{\lambda}(x)$ diverges. This can happen for any estimating prior with a lighter tail than an exponential distribution. We therefore suggest using an exponential distribution for the estimating prior to ensure the loss is not too great. This has the added mathematical convenience that the posterior mean is easily calculated as $\hat{\theta} = x - \lambda\sigma^2$ for some constant λ . If we use an improper exponential prior with density proportional to $e^{-\lambda\theta}$ for all θ (not just $\theta > 0$) then this formula for the posterior mean is exact. Indeed the posterior distribution is given by

$$\pi_x(\theta) \propto e^{-\lambda\theta} e^{-\frac{(\theta-x)^2}{2\sigma^2}} = e^{-\frac{(\theta-x+\lambda\sigma^2)^2}{2\sigma^2} + \frac{\lambda^2\sigma^2}{2} - \lambda x} \propto e^{-\frac{(\theta-x+\lambda\sigma^2)^2}{2\sigma^2}}$$

which is the density of a normal distribution with mean $x - \lambda\sigma^2$ and variance σ^2 .

In this proposition, part (ii) gives the measure of the cost of using the wrong estimating prior. (i) and (iii) give measures of the overall difficulty of the ranking problem. (i) is the irreducible cost of misranking. (iii) is the additional cost from using the point estimates to rank, instead of using the posterior mean. It is an indication of the extent to which the ranking can be improved by using Bayesian methods.

3 Shapes of ranking thresholds

Henderson and Newton (2015) describe different ranking methods in terms of the shapes of what they refer to as “threshold functions”, namely the functions $t_\alpha(\sigma^2)$ which are the smallest value of x , such that the observation (x, σ^2) is ranked in the top α proportion under the ranking method in question. These threshold functions are curves joining

points of equal rank: we will therefore refer these curves as *isotaxes* (singular: *isotaxis*, from Greek *iso* meaning equal, and *taxis* meaning rank). Henderson and Newton (2015) then describe their *r*-values procedure directly by calculating the shape of these isotaxes. We will examine the shape of the isotaxes as a method to better determine the effect of the estimating prior on ranking.

For Bayesian methods, the shape of these isotaxes depends heavily on the choice of estimating prior. For the normal estimating prior with mean 0 and variance τ^2 , for an observation x with standard error σ , the posterior mean is $\frac{\tau^2}{\tau^2 + \sigma^2}x$, so isotaxes are given by solutions to $\frac{\tau^2}{\tau^2 + \sigma^2}x = C$ for constant C , or to $\sigma^2 = \frac{\tau^2}{C}x - \tau^2$. When plotted on a graph of σ^2 against x , these are lines of varying slope, with shallower slope at higher ranks. (Indeed, these lines all pass through the point $(0, -\tau^2)$.)

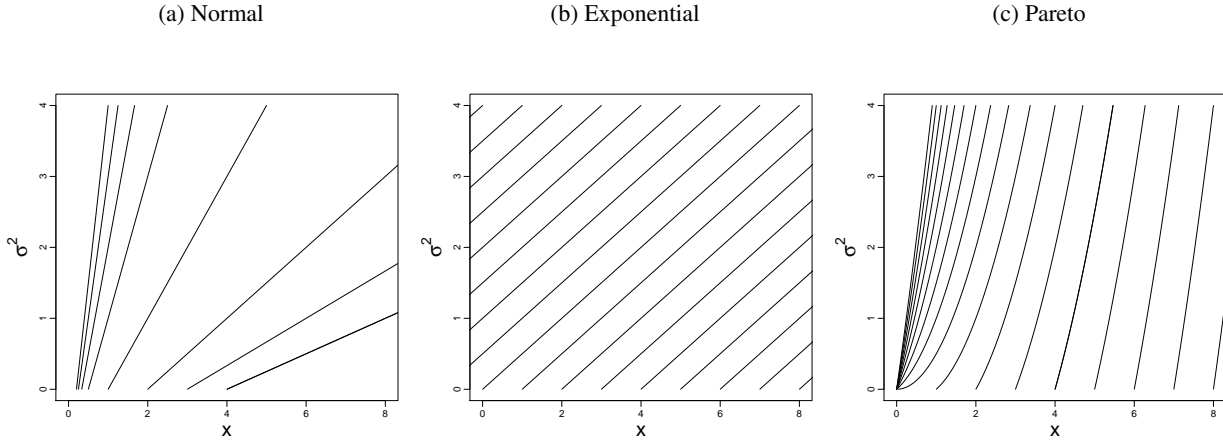
For an exponential estimating prior with hazard rate λ , as mentioned above, the posterior mean is given by $x - \lambda\sigma^2$. The isotaxes are therefore given by the equation $x - \lambda\sigma^2 = C$, or $\sigma^2 = \frac{x}{\lambda} - \frac{C}{\lambda}$, so they are lines of constant slope.

For a heavy-tailed distribution, recall that we have posterior mean approximately $x + \frac{\pi'(x)}{\pi(x)}\sigma^2$. Therefore the isotaxes are functions of the form $x + \frac{\pi'(x)}{\pi(x)}\sigma^2 = C$. A typical example is $\pi(x) = x^{-\alpha}$, so that $\frac{\pi'(x)}{\pi(x)} = -\frac{\alpha}{x}$. This means the isotaxes are curves of the form

$$\begin{aligned} x - \frac{\alpha\sigma^2}{x} &= C \\ x^2 - \alpha\sigma^2 &= Cx \\ \sigma^2 &= \frac{1}{\alpha} \left(x - \frac{C}{2} \right)^2 - \frac{C^2}{4\alpha} \end{aligned}$$

which gives a parabola. We plot the shapes of the isotaxes for these estimating prior distributions in Figure 2.

Figure 2: Isotaxis plots for various choices of estimating prior distribution using posterior mean ranking



We see that for the exponential and heavy-tailed estimating priors, the slopes of isotaxes are bounded away from zero, so the posterior mean cannot be very far from the point estimate for x . Since by assumption, the true value also will not be so far from the point estimate, this means that the posterior mean cannot be too far from the true value.

From the shapes in Figure 2, we see that for the normal estimating prior, the standard error becomes increasingly important as we move towards the tail of the distribution, and that the posterior mean can be arbitrarily far away from

the true value. For the exponential estimating prior, the standard error remains equally important throughout. For the heavy-tailed estimating prior, the standard error becomes less important as we move to the tail of the distribution. Furthermore, the standard error is most important for small standard error, and differences in standard error become less important as the standard error increases.

4 Non-parametric Prior

It is also possible to calculate a non-parametric maximum likelihood estimate for the prior distribution. It was shown by Laird (1978) that the prior in this case is a discrete distribution with finite support. An implementation of this non-parametric prior estimation is given in the `rvalues` package in R. However, this implementation is buggy, so we were unable to compare this method in Section 5. We show that for such a choice of estimating prior, provided the support of the prior distribution includes points sufficiently close to all the observed data, then the posterior mean estimators are robust. Proofs of the following lemmas are in Appendix B.

Lemma 4.1. *Let π be a discrete distribution with probability at least $\frac{1}{r+1}$ in the interval $[x - a, x + a]$ for some $a > 0$. Let $\hat{\theta}$ be the posterior mean for an observation x with standard error σ . Then*

$$|\hat{\theta} - x| \leq a + \sigma \sqrt{2 \log(r)}$$

This means that provided the prior distribution assigns some probability to a region near to each observed value of x , then the posterior mean estimate will have some robustness to model misspecification.

Lemma 4.2. *For a sample of n datapoints and their corresponding standard errors, the non-parametric MLE estimate for the prior distribution always assigns probability at least $\frac{1-e^{-\frac{1}{2}}}{n}$ to the interval $(x - \sigma \sqrt{2 \log(n) + 1}, x + \sigma \sqrt{2 \log(n) + 1})$, for every observed data point (x, σ) .*

From the preceding lemmas, we conclude that ranking based on posterior mean under the non-parametric MLE estimate for the prior is relatively robust, with

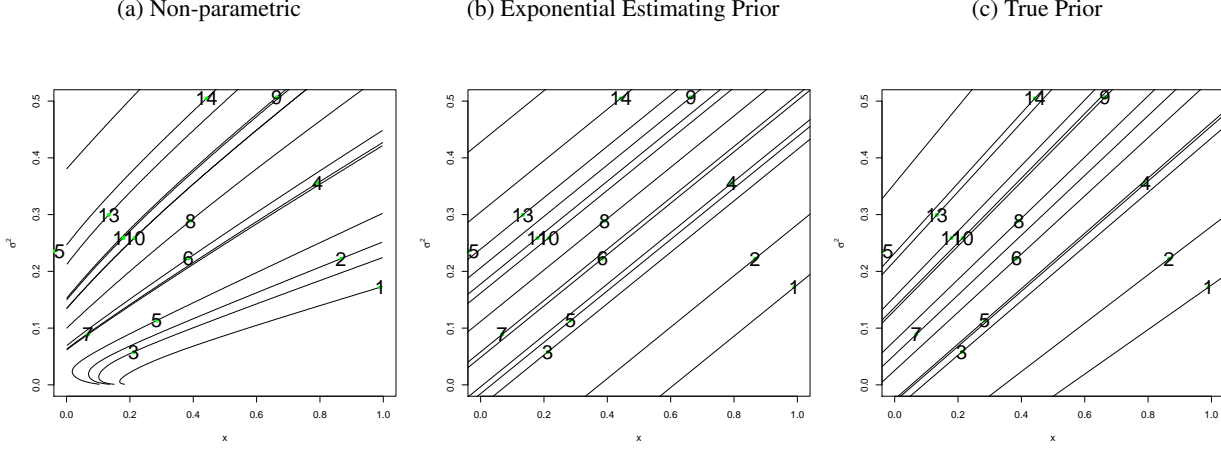
$$|\hat{\theta} - x| \leq \sigma \left(\sqrt{2 \log(n) + 1} + \sqrt{2 \log \left(\frac{n}{(1 - e^{-\frac{1}{2}})} - 1 \right)} \right)$$

We also know that for large n , the non-parametric MLE estimate is consistent, so the ranking will be optimal with the non-parametric MLE. Overall, we conclude that non-parametric estimation of prior provides a reasonable compromise between efficiency and robustness.

However, as is typically the case with non-parametric methods, there is a trade-off between bias and variance. For the non-parametric method, the estimated ranking is asymptotically unbiased, but can have fairly large variance for smaller sample sizes. Figure 3 gives an illustration of this.

We can see that while the non-parametric approach has the isotaxes in approximately the right direction for larger variances, they are somewhat distorted for smaller variances. This is particularly observable at the tail, because the support of the MLE (which is discrete by Laird (1978)) is fairly sparse around the tail. This has a big effect on the posterior mean estimates for points with small standard error. However, it is worth noting that this distortion usually has limited influence on the estimated ranking. The reason for this is that the distortion is only for small standard error, compared with the standard error of the data points, so if some of the data points have small standard error, the isotaxes for posterior mean ranking based on the MLE prior will be close to the correct isotaxes except for very small standard error. Meanwhile, if the standard errors are large, the MLE isotaxes will become further from the

Figure 3: Comparison of Isotaxes for Non-parametric and Parametric Estimation



Isotaxes for the upper tail of simulated data. 500 data points were simulated with the true means following a normal distribution with mean -2.3 and variance 1 . Variances for the observed data points are simulated following a gamma distribution with shape parameter 2 and scale parameter 0.1 . Plot (a) shows the isotaxes for the non-parametric MLE estimate for the prior distribution. Plot (b) shows the isotaxes for an exponential estimating prior. Plot (c) shows the isotaxes for the true prior. Points are numbered according to their rank by posterior means under the true prior. Note that some points are outside the region shown, hence the missing numbers. The isotaxes shown are the ones passing through observed data points.

correct isotaxes, but not many of the observed data points will be included in this region where the isotaxes are far from optimal. The example given in Figure 3 is a typical example where the non-parametric MLE prior gives a poor ranking. There are other typical examples where the MLE prior does not give such a poor ranking.

5 Simulation

5.1 Simulation Design

We use three simulation distributions for the priors (both the true priors and the estimating priors): A normal distribution with known mean 0 and variance τ^2 ; An exponential distribution with hazard rate λ ; and a Pareto distribution with density function $\pi(\theta) = \frac{\alpha\eta^\alpha}{\theta^{\alpha+1}}$ for $\theta > \eta$ where we take $\eta = \frac{1}{2}$ as known. (We have taken one parameter as known for the normal and Pareto distributions, so that each prior has one hyperparameter to be estimated.) For the true priors in the simulation, we set $\tau = 1$ for the normal distribution, $\lambda = 1$ for the exponential distribution and $\alpha = 2$ for the Pareto distribution. For each simulation distribution, we simulate datasets of size 1000 , 10000 , and 100000 .

We simulate the standard error σ for each data set as following an exponential distribution. We present results for the mean of this exponential distribution equal to 0.02 . Results for mean 0.01 , 0.05 and 0.1 are presented in the supplementary materials. The values of σ are independent of the values of θ and values of σ for different data points are independent. To avoid some computational issues caused by values of σ too close to 0 , we added 0.0001 to all values of σ . We do not expect this to significantly impact the results, but we found that some numerical integration routines produced errors when the value of σ was very close to zero.

For each simulated dataset, we analyse with each of the normal, exponential and Pareto distributions as the

estimating prior. We will assess the performance of the ranking by the average increase in the loss function from using the given estimating prior compared to using the true prior. That is, the loss function is:

$$L = \sum_{i=1}^{0.1k} (0.1k - i)(\theta_{(i)} - \theta_{[i]}) \quad (2)$$

where $\theta_{(i)}$ is the true value of θ for the i th ranked unit under the true prior, and $\theta_{[i]}$ is the true value of θ for the i th ranked unit under the estimating prior.

5.2 Theoretical Analysis of Expected Loss for Simulation Distributions

In order to better understand issues related to parameter estimation, we examine the loss function for both optimal parameter estimates (based on minimising the expected loss function) and estimated parameters (estimated from the upper tail of the data). We do not compare the effect of estimating the hyperparameters from the whole data set because two of the distributions used for analysis had support only on the positive real numbers, so estimating these based on the whole data including negative values might lead to strange results. Even if the supports were all the same, estimating the parameters for the estimating prior based on the whole data set when the focus is on the ranking of the top units leads to suboptimal results in ranking.

We calculate the expected loss function in each case (details in Appendix C). Table 1 gives the expected loss function (using Equation (1)) as a function of the true and estimated parameters for each scenario. The optimal parameter values for the estimating priors are therefore the values that minimise these loss functions. Table 2 gives the optimal parameter values in all scenarios, and the corresponding expected additional loss in each scenario from using the misspecified estimating prior distribution. The final column uses the point estimate instead of posterior mean ranking.

From Table 1, We see that the loss functions are quadratic in the parameters of the estimating prior (or in $\frac{1}{\tau^2}$ for the normal distribution). This means that the sensitivity of the loss function to misestimation of the parameter values is roughly proportional to the mean squared difference between the parameter estimate and the optimal value. We calculate the constants of proportionality for our particular choices of parameter values in Table 3. This gives a measure of the sensitivity of the loss function to errors in parameter estimation.

As we see in Table 3, the normal estimating prior is most sensitive to parameter estimation. This makes sense, since the variance of the normal distribution has a very significant impact on the slopes of the isotaxes in the tail of the distribution. The exponential estimating prior is less sensitive to misestimation of parameter values, and the Pareto estimating prior is least sensitive to parameter estimates. This is because in the tail of the distribution, the isotaxes for the Pareto estimating prior become very steep, regardless of the parameter estimates. This indicates an advantage of using a heavy-tailed estimating prior, particularly for small sample sizes, where our parameter estimates have higher MSE. Even for large sample sizes, the parameter estimates are likely to be different from the optimal values, because we typically estimate parameters by a method such as MLE, based on the observed data. We were only able to optimise the loss function for the simulations where we knew the true prior distribution, but in a real situation we would not know the true prior. The parameters estimated by MLE are not optimal for posterior mean ranking.

We now look at the question of parameter estimation. Because we are interested in fitting the tail of the distribution well, we truncate the distribution at the 90th percentile (for the simulations, we used the 90th percentile of the true prior), and estimate the parameters by maximum likelihood for the truncated distribution. Details of the MLE estimates, with derivation, are in Appendix C.3 We compare the theoretically best values and the expected MLE estimates in Table 4. (The Pareto distribution used for simulation has infinite variance, so the MLE estimate for the normal variance does not converge to a constant as sample size increases.) Some of the MLE estimates used here are

Table 1: Expected Loss functions

True	Estimate	Loss function
Normal	Normal	$\left(\frac{1}{\tau^2} - \frac{1}{\hat{\tau}^2}\right)^2 \left(\frac{ae^{-\frac{a^2}{\tau^2}}}{4\pi} + \frac{\tau}{4\sqrt{\pi}} \left(1 - \Phi\left(\frac{\sqrt{2}a}{\tau}\right)\right) \right)$
Normal	Exponential	$\frac{\hat{\lambda}^2}{2\sqrt{\pi}\tau} \left(1 - \Phi\left(\frac{\sqrt{2}a}{\tau}\right)\right) - \frac{\hat{\lambda}e^{-\frac{a^2}{\tau^2}}}{2\pi\tau^2} + \frac{ae^{-\frac{a^2}{\tau^2}}}{4\pi\tau^4} + \frac{1 - \Phi\left(\frac{\sqrt{2}a}{\tau}\right)}{4\sqrt{\pi}\tau^3}$
Normal	Pareto	$\frac{1}{2\pi\tau^2} \left((\hat{\alpha} + 1)^2 \int_a^\infty \frac{1}{\theta^2} e^{-\frac{\theta^2}{\tau^2}} d\theta - \left(2\hat{\alpha} + \frac{3}{2}\right) \frac{\sqrt{\pi}}{\tau} \left(1 - \Phi\left(\frac{\sqrt{2}a}{\tau}\right)\right) + \frac{ae^{-\frac{a^2}{\tau^2}}}{2\tau^2} \right)$
Exponential	Normal	$\frac{e^{-2\lambda a}}{\lambda} \left(\frac{\lambda^4}{2} - \frac{2\lambda^3 a + \lambda^2}{2\hat{\tau}^2} + \frac{2\lambda^2 a^2 + 2\lambda a + 1}{4\hat{\tau}^4} \right)$
Exponential	Exponential	$\frac{(\lambda - \hat{\lambda})^2}{2\lambda} e^{-2\lambda a}$
Exponential	Pareto	$\lambda^2 \left((\hat{\alpha} + 1)^2 \int_a^\infty \frac{1}{\theta^2} e^{-2\lambda\theta} d\theta - 2\lambda(\hat{\alpha} + 1) \int_a^\infty \frac{1}{\theta} e^{-2\lambda\theta} d\theta + \frac{\lambda}{2} e^{-2\lambda a} \right)$
Pareto	Normal	$\frac{\alpha^2 \eta^{2\alpha}}{a^{2\alpha+3}} \left(\frac{(\alpha + 1)^2}{(2\alpha + 3)} - \frac{2(\alpha + 1)a^2}{(2\alpha + 1)\hat{\tau}^2} + \frac{a^4}{(2\alpha - 1)\hat{\tau}^4} \right)$
Pareto	Exponential	$\frac{\alpha^2 \eta^{2\alpha}}{a^{2\alpha+3}} \left(\frac{(\alpha + 1)^2}{(2\alpha + 3)} - \hat{\lambda}a + \frac{\hat{\lambda}^2 a^2}{(2\alpha + 1)} \right)$
Pareto	Pareto	$\frac{\alpha^2 (\alpha - \hat{\alpha})^2 \eta^{2\alpha}}{(2\alpha + 3)a^{2\alpha+3}}$

Table 2: Expected loss — Optimal parameter values (values which minimise the expected loss function from Table 1) and the resulting values of the expected loss function compared with the true prior

(a) Optimal Parameter Values					(b) Expected Loss				
		Estimating Prior					Estimating Prior		
		Normal $\hat{\tau}$	Exp. $\hat{\lambda}$	Pareto $\hat{\alpha}$			Normal	Exp.	Pareto
True Prior	Normal ($\tau = 1$)	1	1.561	1.290	True Prior	Normal ($\tau = 1$)	0	0.00062	0.00208
	Exp. ($\lambda = 1$)	1.701	1	1.677		Exp. ($\lambda = 1$)	0.00015	0	0.00010
	Pareto ($\alpha = 2$)	1.179	1.581	2		Pareto ($\alpha = 2$)	0.00208	0.00036	0
					Point Estimate				

Table 3: Misestimation Loss — The loss functions are all quadratic in the estimated parameter (or $\frac{1}{\tau^2}$ for the normal). This table gives the second derivative of the loss function — it gives an indication of the relative cost of misestimating the parameter values. Optimal parameter estimates are given in Table 2(a). If the optimal parameter value is θ and the value used is θ' , then the additional loss is $a(\theta - \theta')^2$, where a is the number in this table. For the normal estimating prior, the parameter to be estimated is $\frac{1}{\tau^2}$, rather than τ . For the exponential it is the rate λ . For the Pareto, it is the index α . This table gives a measure of the sensitivity of each estimating prior to misestimation of the parameter.

		Estimating Prior		
		Normal	Exp.	Pareto
True Prior	Normal	0.04933429	0.009862926	0.004307
	Exponential	0.04052242	0.005	0.0006835
	Pareto	0.02108185	0.005059644	0.001445613

approximate, so may not exactly reflect parameter values; empirical mean parameter estimates are in Table 6. We see that the expected MLE estimates are in many cases quite far from the optimal values (and the empirical mean for the simulations are also far from optimal). As a consequence, we expect using MLE to estimate hyperparameter values to lead to substantially worse ranking than using the optimal values.

Table 4: Parameter values. Left: optimal parameter values (repeated from Table 2(a)) that minimise expected loss over top 10% of data. Right: expected MLE estimates for parameter values estimated from truncated data.

(a) Optimal Parameter Values					(b) Expected Parameter Estimates				
		Estimating Prior					Estimating Prior		
		Normal	Exp.	Pareto			Normal	Exp.	Pareto
True Prior	Normal	1	1.5614	1.2898	True Prior	Normal	1	2.1122	3.47
	Exponential	1.7005	1	1.6772		Exponential	2.5701	1	3.15
	Pareto	1.1785	1.5811	2		Pareto	NA	0.6325	2

5.3 Simulation Results

The results of the simulation are shown in Table 5. This table gives the average of loss function from Equation (2) over the simulated datasets, for each scenario. As expected, with optimal parameter estimates, using the normal estimating prior when the true prior is heavy-tailed causes a bigger loss, relative to the difficulty of the problem (measured as the loss arising from using a point estimate), than using a heavy-tailed estimating prior when the true prior is normal — when the true prior is normal, the problem is much more difficult (the increase in loss from using the point estimate is larger), but the increase in loss from using the Pareto estimating prior is about the same as the increase when using a normal estimating prior in the easier case where the true prior follows a Pareto distribution. When we use estimated hyperparameter values, the loss from using the Pareto estimating prior when the true prior is normal is larger than using a normal estimating prior when the true prior is Pareto, even taken relative to the loss from using a point estimate. This is explained by the fact that the MLE estimate for the Pareto parameter is further from the optimal value than the MLE estimate of $\frac{1}{\tau^2}$ is from its optimal value. Using an exponential estimating prior does not perform too badly in any of the cases. All methods perform much better than the use of the point estimates. These results show a similar result to the theoretically estimated values in Table 2(b) with many values approximately proportional to that table. The error in the case when the estimating prior is normal and the true prior is heavy-tailed, is theoretically bounded because the Pareto distribution has $\alpha > 0.5$, but results are still poor.

Table 5: Simulation Results: average over simulated data sets of loss function (from Equation 2). Left tables use optimal parameter values. Right tables use MLE estimated parameter values (for data truncated at the true 90th percentile). Top row is for sample size 1000 (1000 datasets), middle row sample size 10000 (100 datasets), bottom row 100000 (10 datasets). Mean of σ is 0.02.

(a) Theoretical, sample size 1000						(b) Estimated, sample size 1000					
		Estimating Prior			Point			Estimating Prior			Point
		Normal	Exp.	Pareto	Estimate			Normal	Exp.	Pareto	Estimate
True Prior	Normal	0	0.003	0.008	0.038	True Prior	Normal	0.000	0.008	0.039	0.038
	Exp.	0	0	0	0.009		Exp.	0.001	0.000	0.003	0.009
	Pareto	0.003	0	0	0.022		Pareto	0.017	0.008	0.000	0.022
(c) Theoretical, sample size 10000						(d) Estimated, sample size 10000					
		Estimating Prior			Point			Estimating Prior			Point
		Normal	Exp.	Pareto	Estimate			Normal	Exp.	Pareto	Estimate
True Prior	Normal	0	0.12	0.35	3.82	True Prior	Normal	0.00	0.61	3.57	3.82
	Exp.	0.01	0	0.02	0.75		Exp.	0.07	0.00	0.25	0.75
	Pareto	0.31	0.05	0	2.03		Pareto	1.74	0.73	0.00	2.03
(e) Theoretical, sample size 100000						(f) Estimated, sample size 100000					
		Estimating Prior			Point			Estimating Prior			Point
		Normal	Exp.	Pareto	Estimate			Normal	Exp.	Pareto	Estimate
True Prior	Normal	0	10.1	32.7	385.1	True Prior	Normal	0.0	59.5	353.6	385.1
	Exp.	2.0	0	2.0	76.5		Exp.	8.2	0.1	27.1	76.5
	Pareto	33.0	5.9	0	209.0		Pareto	187.0	77.8	0.1	209.0

Table 6 gives the mean parameter estimates in cases where we used MLE to estimate parameter values. We see that these are mostly as predicted in Table 4b(b). The main difference is when we use a normal estimating prior generated under an exponential true prior. Here the estimated value is much closer to the optimal value. This is because the approximation we used in deriving the expected MLE estimate is not very accurate. This explains why the normal estimating prior with estimated parameter values did not perform so poorly in this scenario. We know that the ranking based on a normal estimating prior is most sensitive to parameter estimates. However, because MLE provides a fairly good estimate in this case, the loss from using an estimated value is not so great. For the exponential and Pareto estimating priors, the MLE does not provide good parameter estimates for the purpose of ranking. Because the ranking loss in these cases is less sensitive to estimation errors in the hyperparameters, the resulting losses are not excessive. However, this indicates there is great scope for improving results by devising better parameter estimation techniques. It is also worth noting that these hyperparameters were estimated to fit the tail well, rather than the whole dataset. More common practice is to estimate the hyperparameters based on the whole data. We would expect this to result in much worse ranking results, particularly for the normal estimating prior where the loss is particularly sensitive to the parameter estimates.

Table 6: parameter estimates

True Prior	Sample size	Normal	Exponential	Pareto
Normal	1000	0.999(0.0883)	2.135(0.1902)	3.479(0.2631)
	10000	1.000(0.0186)	2.120(0.0586)	3.462(0.0811)
	100000	1.000(0.0061)	2.119(0.0192)	3.460(0.0265)
Exponential	1000	2.051(0.1539)	1.012(0.1031)	3.113(0.2538)
	10000	2.055(0.0485)	1.002(0.0316)	3.094(0.0780)
	100000	2.055(0.0150)	1.001(0.0097)	3.093(0.0236)
Pareto	1000	36.081(232.50)	0.674(0.1494)	2.024(0.2059)
	10000	80.509(445.18)	0.639(0.0589)	2.005(0.0640)
	100000	99.292(487.59)	0.634(0.0225)	2.003(0.0200)

6 Real Data Analysis

6.1 Type 2 Diabetes

We look at several real data sets. These datasets were studied by Henderson & Newton (2015) for their work on r -values. The first data set consists of GWAS data for log odds ratio between SNPs and type-2 diabetes from Morris *et al.* (2006). The data are available from <http://diagram-consortium.org/downloads.html>. The data consists of 137,899 SNPs from 12,171 type 2 diabetes cases and 56,862 controls. For each SNP, an odds ratio is available along with a 95% confidence interval. Following Henderson & Newton (2015), we have taken the value as the log-odds ratio, assuming this estimate follows a normal distribution, and that the standard deviation of this distribution is one-quarter of the width of the log of the 95% confidence interval provided. The resulting positive data points and isotaxes for a normal and exponential estimating prior are shown in Figure 4.

From this figure, we see that using a normal estimating prior with naively estimated variance, the estimated variance is small, leading to isotaxes passing close to the origin. This makes the ranking focus on values with small variance, and rank values with larger observed value and larger variance behind values with smaller variance. The exponential estimating prior provides a ranking that selects many more of the points with large estimated effect size. We can improve the performance of a normal prior by artificially inflating the variance to match the tail better.

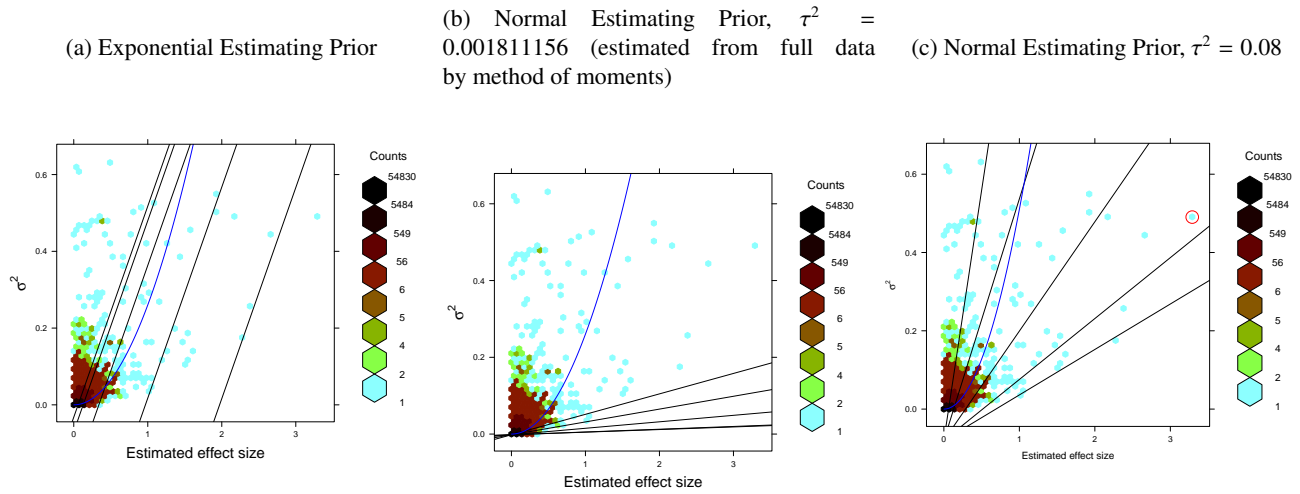


Figure 4: Isotaxes for exponential and normal estimating prior distributions for type-2 diabetes data. Isotaxes shown are for top 5%, top 1%, top 0.1%, top 0.01% and top 0.001%. The blue curve represents the 95% significance level against the null hypothesis — that is, points to the left of the blue curve do not have estimated effect size significantly different from zero.

Figure 4c shows the effect of this. This does select a lot more of the points with large effect sizes, and indeed the top 1% isotaxis is very similar to the 1% isotaxis for the exponential estimating prior. On the other hand, the higher isotaxes do put too much weight on having smaller standard error, ranking a number of points with smaller estimated effect size ahead of the point with largest estimated effect size (circled in red). It is extremely implausible that this ranking is correct. Overall, the ranking based on an exponential estimating prior appears more plausible to us for this dataset.

6.2 Breast Cancer

Next we look at the gene expression data relating to breast cancer from West *et al.* (2001). This dataset is available in the *rvalues* package. The data set consists of gene expression measurements of 7129 genes across 49 breast tumour samples — 25 oestrogen receptor (ER)+ samples and 24 ER− samples. For each gene, the difference in means between the ER+ and ER− groups is calculated, along with its appropriate standard error. In theory the error distribution should be modelled as a *t*-distribution with 47 degrees of freedom. However, for the purpose of this paper, we have used a normal distribution. The loss of accuracy should be fairly small. The resulting plot of variance against estimated effect size, and isotaxes for posterior mean with an exponential and a normal estimating prior are shown in Figure 5.

As for the diabetes data, we see that the normal estimating prior results in very flat isotaxes, and therefore gives a high ranking to observations with small variance. Meanwhile, the exponential estimating prior puts a lot more weight on points with large estimated effect size. Again, we see that using a normal estimating prior with inflated variance results in isotaxes that are more similar to the exponential estimating prior. In this case, the differences between the two rankings are not so clear-cut as the previous case, where some of the rankings using a normal estimating prior with increased variance were completely implausible. In this case, both the rankings for the exponential estimating

(a) Exponential Estimating Prior

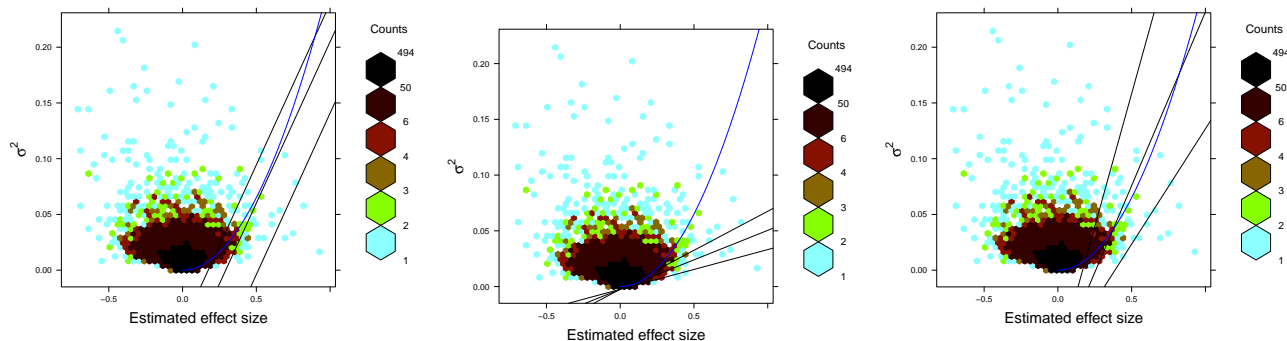
(b) Normal Estimating Prior, $\tau^2 = 0.002478642$, estimated from data by method of moments(c) Normal Estimating Prior, $\tau^2 = 0.08$ 

Figure 5: Isotaxes for exponential and normal estimating prior distributions for breast cancer data. Isotaxes shown are for top 5%, top 1%, and top 0.1%. The blue curve represents the 95% significance level against the null hypothesis — that is, points to the left of the blue curve do not have estimated effect size significantly more than zero using a two-sided test.

prior and the normal estimating prior with inflated variance seem reasonable.

7 Conclusions and Future Work

We have seen that choice of estimating prior can have a very large effect on Bayesian ranking methods. For the majority of ranking problems, we are particularly interested in ranking at the upper tail of the distribution. The ranking of the upper tail can be particularly affected by choice of estimating prior.

Using a light-tailed estimating prior for posterior mean ranking can lead to very bad results. If the true prior is heavy-tailed, the posterior mean can be far away from the truth. Conversely, if the estimating prior is too heavy-tailed, the posterior mean estimated will be between the posterior mean under the true prior and the point estimate. This cannot be too far from the true posterior mean. This means that using an exponential or heavier-tailed distribution as the estimating prior should be more robust to model misspecification.

In addition to being less robust to model misspecification, light-tailed estimating priors can be more sensitive to estimated parameter values. In cases where the estimating prior is misspecified, using MLE estimates for hyperparameters can also be far from optimal, so this can lead to bad results even in cases with large datasets. Since we are usually particularly interested in the top units, it is usually advisable to choose parameter values that fit the tail of the distribution well.

Using a non-parametric prior is robust, in that there is an upper bound on how far the posterior mean can be from the posterior mean under the true prior. However, using a non-parametric prior can be inefficient for smaller sample sizes, and can lead to some strange rankings.

We have confirmed our results by simulation studies and real data examples. In the simulation study, we found that an exponential estimating prior performed relatively well regardless of the true prior. Our simulation study

also studied the effect of estimating hyperparameters on the performance. As expected, estimating hyperparameters does cause some loss. The estimation in this simulation was done by maximum likelihood. However, since the loss function we are aiming to minimise is not the standard squared error loss, this is not the optimal estimation method. We based our parameter estimation on the upper 10% of the data points. It is common for analyses which use the whole data to estimate hyperparameters. Doing this could lead to far worse results when the estimating prior is misspecified.

Overall, unless there is good evidence otherwise, we suggest an exponential estimating prior will be a good compromise between robustness and efficiency in most cases. It also offers easy computation of posterior mean.

7.1 Future Work

The most obvious direction to need improvement in this research is hyperparameter estimation. We have seen in our simulation study that estimation by MLE can lead to bad ranking results. This is because the loss function from prior misspecification is different from the loss function that MLE estimation aims to minimise. This suggests that a different method of estimating the hyperparameters is needed — a method specifically targeted at optimising ranking estimates. We know the loss function that we are aiming to minimise, so it should be possible to find an explicit way to solve this and derive a procedure for estimating the hyperparameter. Given our recommendation to use an exponential estimating prior in most cases, finding the best hyperparameters should not prove too challenging a problem.

Our study has a number of limitations. We have considered only cases where ranking is by posterior mean and the error distribution is normal. In future work, we should study the problem for different error distributions, not just normal. Further estimation is also needed into cases where the variance of the error distribution depends on the parameter θ . This can allow certain approximations to be applied. For example a Poisson distribution can be approximated by a normal distribution where the variance depends on the mean. We should also study the problem for different methods and objective functions, e.g. r -values, posterior expected rank.

We also have not considered the effect of model selection on ranking. If the estimating prior distribution is chosen based on certain model selection criteria, this may improve the ranking. However, model selection for mixture models can be difficult, so it might not provide the improvements we hope for. Model selection also depends upon a good set of candidate models. Our research suggests that the form of the function $\lambda(x) = -\frac{\pi'(x)}{\pi(x)}$ is most crucial in our choice of estimating prior, so including a sufficient range of models to allow flexibility in this function should allow us to obtain good ranking results, provided the model selection criteria are well chosen to be related to our objective function.

A Loss Function Calculation

Proposition A.1. *Suppose the true prior distribution of the parameter θ has density function $\pi(\theta)$, and that we have two observations x_1 and x_2 which are normally distributed with means θ_1 and θ_2 and standard deviations σ_1 and σ_2 respectively, where θ_1 and θ_2 are random samples from the true prior distribution, and σ_1 and σ_2 are assumed to be small.*

- (i) *The expected loss when the estimating prior and the true prior are the same (which we will refer to as the optimal expected loss) is approximately given by*

$$\frac{\sigma_1^2 + \sigma_2^2}{2} \mathbb{E}(\pi(x))$$

(ii) When the estimating prior has density $\hat{\pi}$, the difference between the expected loss and the optimal expected loss is approximately given by

$$\frac{1}{2}(\sigma_1^2 - \sigma_2^2)^2 \int_a^\infty \pi(x)^2 (\lambda(x) - \hat{\lambda}(x))^2 dx$$

where $\hat{\lambda}(x) = -\frac{\hat{\pi}'(x)}{\hat{\pi}(x)}$.

(iii) The difference between the expected loss from using the point estimate x_i and the optimal expected loss is approximately given by

$$\frac{1}{2}(\sigma_1^2 - \sigma_2^2)^2 \int_a^\infty \pi'(x)^2 dx$$

Proof. (i) Suppose that the true parameter values are θ_1 and θ_2 respectively. Let $\Delta = \theta_1 - \theta_2$. Now the loss from mis-ranking is $|\Delta|$ if the points are mis-ranked and 0 if they are not misranked. The points are misranked if either $\Delta > 0$ and $x_1 - \lambda(x_1)\sigma_1^2 < x_2 - \lambda(x_2)\sigma_2^2$ or if $\Delta < 0$ and $x_1 - \lambda(x_1)\sigma_1^2 > x_2 - \lambda(x_2)\sigma_2^2$. Since the points will not plausibly be misranked if Δ is large (since x_1 and x_2 will then with high probability be far apart), we will assume that Δ is small, so that we have $\lambda(x_1) \approx \lambda(x_2) \approx \lambda(\theta_1)$. We will denote this common value λ . Now for fixed θ_1 and θ_2 , suppose $\Delta > 0$; we want to calculate $P(x_1 - \lambda\sigma_1^2 < x_2 - \lambda\sigma_2^2)$. We know that $x_1 - x_2$ is normally distributed with mean Δ and variance $\sigma_1^2 + \sigma_2^2$. Therefore,

$$P(x_1 - \lambda\sigma_1^2 < x_2 - \lambda\sigma_2^2) = \Phi\left(\frac{\lambda(\sigma_1^2 - \sigma_2^2) - \Delta}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)$$

On the other hand, if $\Delta < 0$, we have

$$P(x_1 - \lambda\sigma_1^2 > x_2 - \lambda\sigma_2^2) = \Phi\left(\frac{-\lambda(\sigma_1^2 - \sigma_2^2) + \Delta}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)$$

Now suppose that we fix θ_1 , and we want to take the expected loss over the distribution of θ_2 . This is given by

$$l(\theta_1) = \int_0^\infty \pi(\theta_1 - \Delta) \Delta \Phi\left(\frac{\lambda(\sigma_1^2 - \sigma_2^2) - \Delta}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) d\Delta - \int_{-\infty}^0 \pi(\theta_1 - \Delta) \Delta \Phi\left(\frac{-\lambda(\sigma_1^2 - \sigma_2^2) + \Delta}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) d\Delta$$

Since the probability of misranking is negligible for large Δ , we will consider only small values of Δ . For these small values, we can take the Taylor expansion

$$\pi(\theta_1 - \Delta) = \pi(\theta_1) - \Delta\pi'(\theta_1) = \pi(\theta_1)(1 + \Delta\lambda(\theta_1))$$

Substituting this into the above loss function gives

$$l(\theta_1) = \int_0^\infty \pi(\theta_1)(1 + \Delta\lambda(\theta_1)) \Delta \Phi\left(\frac{\lambda(\sigma_1^2 - \sigma_2^2) - \Delta}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) d\Delta - \int_{-\infty}^0 \pi(\theta_1)(1 + \Delta\lambda(\theta_1)) \Delta \Phi\left(\frac{-\lambda(\sigma_1^2 - \sigma_2^2) + \Delta}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) d\Delta$$

We recall that

$$\begin{aligned}
\int_{-c}^{\infty} \frac{\xi^3 e^{-\frac{\xi^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} d\xi &= \left[-\frac{\sigma}{\sqrt{2\pi}} \xi^2 e^{-\frac{\xi^2}{2\sigma^2}} \right]_{-c}^{\infty} + \int_{-c}^{\infty} \frac{2\sigma}{\sqrt{2\pi}} \xi e^{-\frac{\xi^2}{2\sigma^2}} d\xi \\
&= \frac{1}{\sqrt{2\pi}} (\sigma c^2 + 2\sigma^3) e^{-\frac{c^2}{2\sigma^2}} \\
\int_{-c}^{\infty} \frac{\xi^2 e^{-\frac{\xi^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} d\xi &= \left[-\frac{\sigma}{\sqrt{2\pi}} \xi e^{-\frac{\xi^2}{2\sigma^2}} \right]_{-c}^{\infty} + \int_{-c}^{\infty} \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2\sigma^2}} d\xi \\
&= \sigma^2 \Phi\left(\frac{c}{\sigma}\right) - \frac{\sigma c e^{-\frac{c^2}{2\sigma^2}}}{\sqrt{2\pi}} \\
\int_{-c}^{\infty} \frac{\xi e^{-\frac{\xi^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} d\xi &= \frac{\sigma e^{-\frac{c^2}{2\sigma^2}}}{\sqrt{2\pi}} \\
\int_{-c}^{\infty} \frac{e^{-\frac{\xi^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} d\xi &= \Phi\left(\frac{c}{\sigma}\right)
\end{aligned}$$

Hence we calculate

$$\begin{aligned}
\int_0^\infty \Delta^2 \Phi\left(\frac{c-\Delta}{\sigma}\right) d\Delta &= \left[\frac{\Delta^3}{3} \Phi\left(\frac{c-\Delta}{\sigma}\right) \right]_0^\infty + \int_0^\infty \frac{\Delta^3}{3} \frac{e^{-\frac{(c-\Delta)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} d\Delta \\
&= \int_{-c}^\infty \frac{(\xi+c)^3}{3} \frac{e^{-\frac{\xi^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} d\xi \\
&= \frac{1}{3} \left(\frac{1}{\sqrt{2\pi}} (\sigma c^2 + 2\sigma^3) e^{-\frac{c^2}{2\sigma^2}} + 3c \left(\sigma^2 \Phi\left(\frac{c}{\sigma}\right) - \frac{\sigma c e^{-\frac{c^2}{2\sigma^2}}}{\sqrt{2\pi}} \right) + 3c^2 \frac{\sigma e^{-\frac{c^2}{2\sigma^2}}}{\sqrt{2\pi}} + c^3 \Phi\left(\frac{c}{\sigma}\right) \right) \\
&= \frac{1}{3} \left((\sigma c^2 + 2\sigma^3 - 3c^2 \sigma + 3c^2 \sigma) \frac{e^{-\frac{c^2}{2\sigma^2}}}{\sqrt{2\pi}} + (3c\sigma^2 + c^3) \Phi\left(\frac{c}{\sigma}\right) \right) \\
&= \frac{1}{3} \left((2\sigma^3 + \sigma c^2) \frac{e^{-\frac{c^2}{2\sigma^2}}}{\sqrt{2\pi}} + (3c\sigma^2 + c^3) \Phi\left(\frac{c}{\sigma}\right) \right) \\
\int_0^\infty \Delta \Phi\left(\frac{c-\Delta}{\sigma}\right) d\Delta &= \left[\frac{\Delta^2}{2} \Phi\left(\frac{c-\Delta}{\sigma}\right) \right]_0^\infty + \int_0^\infty \frac{\Delta^2}{2} \frac{e^{-\frac{(c-\Delta)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} d\Delta \\
&= \int_{-c}^\infty \frac{(\xi+c)^2}{2} \frac{e^{-\frac{\xi^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} d\xi \\
&= \frac{1}{2} \left(\left(\sigma^2 \Phi\left(\frac{c}{\sigma}\right) - \frac{\sigma c e^{-\frac{c^2}{2\sigma^2}}}{\sqrt{2\pi}} \right) + 2c \frac{\sigma e^{-\frac{c^2}{2\sigma^2}}}{\sqrt{2\pi}} + c^2 \Phi\left(\frac{c}{\sigma}\right) \right) \\
&= \frac{1}{2} \left((\sigma^2 + c^2) \Phi\left(\frac{c}{\sigma}\right) + \frac{\sigma c e^{-\frac{c^2}{2\sigma^2}}}{\sqrt{2\pi}} \right)
\end{aligned}$$

In the loss function, we let $c = \lambda(\sigma_1^2 - \sigma_2^2)$ and $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$. Substituting these into the loss function gives:

$$\begin{aligned}
l(\theta_1) &= \pi(\theta_1) \left(\int_0^\infty (\Delta + \Delta^2 \lambda) \Phi\left(\frac{c - \Delta}{\sigma}\right) d\Delta - \int_{-\infty}^0 (\Delta + \Delta^2 \lambda) \Phi\left(\frac{-c + \Delta}{\sigma}\right) d\Delta \right) \\
&= \pi(\theta_1) \left(\int_0^\infty (\Delta + \Delta^2 \lambda) \Phi\left(\frac{c - \Delta}{\sigma}\right) d\Delta + \int_0^\infty (\Delta - \Delta^2 \lambda) \Phi\left(\frac{-c - \Delta}{\sigma}\right) d\Delta \right) \\
&= \pi(\theta_1) \left(\int_0^\infty \Delta \Phi\left(\frac{c - \Delta}{\sigma}\right) d\Delta + \int_0^\infty \Delta \Phi\left(\frac{-c - \Delta}{\sigma}\right) d\Delta + \lambda \int_0^\infty \Delta^2 \Phi\left(\frac{c - \Delta}{\sigma}\right) d\Delta - \lambda \int_0^\infty \Delta^2 \Phi\left(\frac{-c - \Delta}{\sigma}\right) d\Delta \right) \\
&= \pi(\theta_1) \left(\frac{1}{2} \left((\sigma^2 + c^2) \left(\Phi\left(\frac{c}{\sigma}\right) + \Phi\left(\frac{-c}{\sigma}\right) \right) + \frac{\sigma e^{-\frac{c^2}{2\sigma^2}}}{\sqrt{2\pi}} (c + (-c)) \right) + \frac{\lambda}{3} \left((3c\sigma^2 + c^3) \Phi\left(\frac{c}{\sigma}\right) + (3c\sigma^2 + c^3) \Phi\left(\frac{-c}{\sigma}\right) \right) \right) \\
&= \pi(\theta_1) \left(\frac{1}{2} ((\sigma^2 + c^2)) + \frac{\lambda}{3} ((3c\sigma^2 + c^3) \left(\Phi\left(\frac{c}{\sigma}\right) + \Phi\left(-\frac{c}{\sigma}\right) \right)) \right) \\
&= \pi(\theta_1) \left(\frac{1}{2} ((\sigma^2 + c^2)) + \frac{\lambda(3c\sigma^2 + c^3)}{3} \right)
\end{aligned}$$

If we let $d = \sigma_1^2 - \sigma_2^2$, so that $c = \lambda d$, then we have

$$l(\theta_1) = \pi(\theta_1) \left(\frac{1}{2} ((\sigma^2 + d^2 \lambda^2)) + \frac{\lambda(3d\lambda\sigma^2 + d^3 \lambda^3)}{3} \right)$$

(For anyone thinking at this point that the dimensions do not work in this formula, it is worthwhile to remember that λ and π are inversely proportional to changes in the scale of θ . That is, if we change the units so that the value of θ doubles, the values of λ and π will be halved.)

With $\frac{d\lambda}{\sigma}$ assumed to be small, we can neglect the $d^3 \lambda^4$ term to get

$$l(\theta_1) = \pi(\theta_1) \left(\frac{1}{2} (\sigma^2 + d^2 \lambda^2) + d\lambda^2 \sigma^2 \right)$$

If we assume that d^2 is negligible, then our expression becomes

$$l(\theta_1) = \pi(\theta_1) \sigma^2 \left(\frac{1}{2} + d\lambda^2 \right)$$

We take the expectation of this over the distribution of θ_1 to get that the expected loss is approximately

$$\frac{\sigma^2}{2} \left(\mathbb{E}(\pi(x)) - 2d\mathbb{E}(\pi(x)\lambda(x)^2) \right)$$

We have assumed that d is small with respect to this second term, so the expected loss is approximately $\frac{\sigma^2}{2} \mathbb{E}(\pi(x))$.

(ii) Since σ_1 and σ_2 are small, we can assume that $\lambda(x_1) \approx \lambda(x_2)$. We will let λ denote this common value. Suppose that x_1 is the observed value of X_1 and x_2 is the observed value of X_2 , and that $x_1 - x_2 > \lambda(\sigma_1^2 - \sigma_2^2)$, but $x_1 - x_2 < \hat{\lambda}(\sigma_1^2 - \sigma_2^2)$ so that x_1 and x_2 are mis-ranked compared to the ranking under the true prior. Let θ be the underlying parameter value for x_1 , and let ϕ be the underlying parameter value for x_2 . The expected increase in the loss function due to this misranking, compared to using the true prior, is then

$$\begin{aligned} \iint \pi_{x_1}(\theta) \pi_{x_2}(\phi) (\theta - \phi) d\theta d\phi &= \int_{-\infty}^{\infty} \pi_{x_1}(\theta) \theta d\theta - \int_{-\infty}^{\infty} \pi_{x_2}(\phi) \phi d\phi \\ &= x_1 - \lambda\sigma_1^2 - (x_2 - \lambda\sigma_2^2) \\ &= x_1 - x_2 - \lambda(\sigma_1^2 - \sigma_2^2) \end{aligned}$$

(where $\pi_{x_1}(\theta)$ and $\pi_{x_2}(\phi)$ are the posterior distributions of θ and ϕ given observations x_1 and x_2 respectively, under the true prior). Similarly, if $\hat{\lambda}(\sigma_1^2 - \sigma_2^2) < x_1 - x_2 < \lambda(\sigma_1^2 - \sigma_2^2)$, then the expected increase in loss is $x_2 - x_1 - \lambda(\sigma_2^2 - \sigma_1^2)$.

Now suppose we fix $X_1 = x_1$ and take the expectation of the loss over X_2 . The expected loss due to mis-ranking them is

$$\int_{x_1 - \lambda(\sigma_1^2 - \sigma_2^2)}^{x_1 - \hat{\lambda}(\sigma_1^2 - \sigma_2^2)} \pi_2(x_2) (x_1 - x_2 - \lambda(\sigma_1^2 - \sigma_2^2)) dx_2 \quad (3)$$

where π_2 is the marginal density of x_2 . In the case where $(\hat{\lambda} - \lambda)(\sigma_1^2 - \sigma_2^2) < 0$, we get this by calculating expected misranking loss over all values for which x_1 and x_2 are misranked (compared to posterior mean ranking using the true prior). In the case where $(\hat{\lambda} - \lambda)(\sigma_1^2 - \sigma_2^2) > 0$, calculating the expected misranking loss over all values where x_1 and x_2 are misranked gives

$$\int_{x_1 - \lambda(\sigma_1^2 - \sigma_2^2)}^{x_1 - \hat{\lambda}(\sigma_1^2 - \sigma_2^2)} \pi_2(x_2) (x_2 - x_1 - \lambda(\sigma_2^2 - \sigma_1^2)) dx_2$$

and by reversing the limits and negating the integrand, we get the formula from Equation 3 in this case also.

Since σ_1^2 and σ_2^2 are both small, we can assume that π, π_1 and π_2 are approximately constant around x_1 , so that $\pi_2(x_2) \approx \pi_1(x_1) \approx \pi(x_1)$ for all x_2 in the relevant range. The integral is then approximately

$$\begin{aligned} \pi(x_1) \int_{x_1 - \lambda(\sigma_1^2 - \sigma_2^2)}^{x_1 - \hat{\lambda}(\sigma_1^2 - \sigma_2^2)} (x_2 - (x_1 - \lambda(\sigma_1^2 - \sigma_2^2))) dx_2 &= \frac{1}{2} \pi(x_1) ((x_1 - \hat{\lambda}(\sigma_1^2 - \sigma_2^2)) - (x_1 - \lambda(\sigma_1^2 - \sigma_2^2)))^2 \\ &= \frac{1}{2} \pi(x_1) (\lambda - \hat{\lambda})^2 (\sigma_1^2 - \sigma_2^2)^2 \end{aligned}$$

For the overall mis-ranking loss, we take the expectation of this over x_1 . We are usually particularly interested in the mis-ranking loss of the upper tail, that is the expected loss due to all misrankings in the upper tail, so we usually take the expectation over the distribution of x_1 for values $x_1 > a$ for some chosen a . This is given by

$$\frac{1}{2} (\sigma_1^2 - \sigma_2^2)^2 \int_a^{\infty} \pi(x_1)^2 (\lambda(x_1) - \hat{\lambda}(x_1))^2 dx_1$$

(iii) we calculate this loss by substituting $\hat{\lambda} = 0$ into our expression for the additional loss, we get

$$\frac{1}{2} (\sigma_1^2 - \sigma_2^2) \int_a^{\infty} \pi(x)^2 \lambda(\theta)^2 d\theta = \frac{1}{2} (\sigma_1^2 - \sigma_2^2) \int_a^{\infty} \pi'(x)^2 dx$$

□

B Non-Parametric Prior

Lemma B.1. Let π be a discrete distribution with probability at least $\frac{1}{r+1}$ in the interval $[x-a, x+a]$ for some $a > 0$. Let $\hat{\theta}$ be the posterior mean for an observation x with standard error σ . Then

$$|\hat{\theta} - x| \leq a + \sigma \sqrt{2 \log(r)}$$

Proof. Let the support of π be the values $x + b_i$, with probabilities π_i . Then for the posterior distribution of θ , the probability of $x + b_i$ is

$$\frac{\pi_i e^{-\frac{b_i^2}{2\sigma^2}}}{\sum \pi_j e^{-\frac{b_j^2}{2\sigma^2}}}$$

The posterior mean is therefore

$$\hat{\theta} = x + \frac{\sum b_i \pi_i e^{-\frac{b_i^2}{2\sigma^2}}}{\sum \pi_j e^{-\frac{b_j^2}{2\sigma^2}}}$$

We see that the difference $|\hat{\theta} - x|$ is maximised when the b_i all have the same sign, which we will w.l.o.g. assume to be positive. It is clear that $\hat{\theta}$ is maximised by setting all the b_i in the interval $[0, a]$ to equal a , since this both minimises the posterior probability of the interval $[x, x+a]$ and maximises the posterior mean conditional on lying in this interval. We will therefore assume that $b_1 = a$, and $\pi_1 = \frac{1}{r+1}$, then we have

$$|\hat{\theta} - x| = \frac{\sum b_i \pi_i e^{-\frac{b_i^2}{2\sigma^2}}}{\sum \pi_j e^{-\frac{b_j^2}{2\sigma^2}}} = \frac{\frac{ae^{-\frac{a^2}{2\sigma^2}}}{r+1} + \sum_{i=2}^k b_i \pi_i e^{-\frac{b_i^2}{2\sigma^2}}}{\sum_{i=1}^k b_i \pi_i e^{-\frac{b_i^2}{2\sigma^2}}} = \frac{ae^{-\frac{a^2}{2\sigma^2}}}{(r+1) \sum \pi_j e^{-\frac{b_j^2}{2\sigma^2}}} + \frac{\sum_{i=m+1}^k b_i \pi_i e^{-\frac{b_i^2}{2\sigma^2}}}{\sum \pi_j e^{-\frac{b_j^2}{2\sigma^2}}}$$

For fixed b_i , and fixed $\sum_{i=2}^k \pi_i$, if $\frac{ae^{-\frac{a^2}{2\sigma^2}}}{(r+1) \sum \pi_j e^{-\frac{b_j^2}{2\sigma^2}}} + \frac{\sum_{i=2}^k b_i \pi_i e^{-\frac{b_i^2}{2\sigma^2}}}{\sum \pi_j e^{-\frac{b_j^2}{2\sigma^2}}} = C$ then $\frac{ae^{-\frac{a^2}{2\sigma^2}}}{(r+1)} + \sum_{i=2}^k b_i \pi_i e^{-\frac{b_i^2}{2\sigma^2}} = C \sum \pi_j e^{-\frac{b_j^2}{2\sigma^2}}$

$$\frac{(a-C)}{(r+1)} e^{-\frac{a^2}{2\sigma^2}} + \sum_{i=2}^k (b_i - C) \pi_i e^{-\frac{b_i^2}{2\sigma^2}} = 0$$

This gives that the contours are linear functions in π , so the maximum value of C occurs at a vertex with only one non-zero value of π_i . The value of $\hat{\theta} - x$ is maximised subject to $b_1 = a$, $\pi_1 = \frac{1}{r+1}$ by setting $k = 2$, $\pi_2 = \frac{r}{1+r}$ and choosing the value of b_2 to maximise the resulting quantity. In this case, we have

$$|\hat{\theta} - x| = \frac{ae^{-\frac{a^2}{2\sigma^2}} + rbe^{-\frac{b^2}{2\sigma^2}}}{e^{-\frac{a^2}{2\sigma^2}} + re^{-\frac{b^2}{2\sigma^2}}} = a + \frac{r(b-a)e^{-\frac{b^2}{2\sigma^2}}}{e^{-\frac{a^2}{2\sigma^2}} + re^{-\frac{b^2}{2\sigma^2}}} = a + \frac{r(b-a)}{r + e^{\frac{b^2-a^2}{2\sigma^2}}}$$

Substituting $a = v\sigma$, $b - a = w\sigma$, this expression becomes

$$\hat{\theta} - x = a + \sigma \frac{rw}{r + e^{\frac{w^2+2vw}{2}}}$$

The derivative of $\frac{r+e^{\frac{w^2+2vw}{2}}}{rw}$ with respect to w is $\left(\frac{w+v}{rw} - \frac{1}{rw^2}\right) e^{\frac{w^2+2vw}{2}} - \frac{1}{w^2}$. We see that for $w > \sqrt{2 \log(r)}$, $w > \sqrt{2}$ and $v \geq 0$, we have

$$\left(\frac{w+v}{rw} - \frac{1}{rw^2}\right) e^{\frac{w^2+2vw}{2}} - \frac{1}{w^2} \geq \left(\frac{w+v}{rw} - \frac{1}{rw^2}\right) e^{\log(r)} e^{vw} - 1 \geq 1 + \frac{v}{w} - \frac{2}{w^2} \geq 0$$

so $\frac{r+e^{\frac{w^2+2vw}{2}}}{rw}$ is increasing. Therefore $\frac{rw}{r+e^{\frac{w^2+2vw}{2}}}$ is decreasing. This means that assuming $r > e$, we have that $\frac{rw}{r+e^{\frac{w^2+2vw}{2}}} \leq \frac{r\sqrt{2\log(r)}}{r(1+e^{\sqrt{2\log(r)}})} \leq \sqrt{2\log(r)}$ for all $w > \sqrt{2\log(r)}$. Meanwhile, we always have $\frac{rw}{r+e^{\frac{w^2+2vw}{2}}} \leq w$, so we always have $\frac{rw}{r+e^{\frac{w^2+2vw}{2}}} \leq \sqrt{2\log(r)}$, and therefore

$$|\hat{\theta} - x| \leq |a| + \sigma \sqrt{2\log(r)}$$

□

Lemma B.2. *For a sample of n datapoints and their corresponding standard errors, the MLE estimate for the prior distribution always assigns probability at least $\frac{1-e^{-\frac{1}{n}}}{n}$ to the interval $(x - \sigma\sqrt{2\log(n)+1}, x + \sigma\sqrt{2\log(n)+1})$, for every observed data point (x, σ) .*

Proof. Suppose the MLE assigns probability π_i to point b_i . We will separate the points b_i into points that are in the interval $I = (x - \sigma\sqrt{2\log(n)+1}, x + \sigma\sqrt{2\log(n)+1})$, and points that are not. Suppose the first m points are in the interval I and the remaining points are not. We are aiming to show that $\sum_{i=1}^m \pi_i > \frac{1}{n}$. Suppose this is not the case. We will then show that the distribution assigning probability π_i to each point b_i is not the MLE by constructing a prior distribution with larger likelihood. Let $\phi = \sum_{i=1}^m \pi_i$. Let \mathcal{X} be the data set, and let $(x, \sigma) \in \mathcal{X}$ be a data point. The log-likelihood of the data can be represented as $l(\mathcal{X} \setminus (x, \sigma)) + l(x, \sigma)$, i.e. as the likelihood of the point (x, σ) plus the likelihood of the remainder of the data points. For a data point (y, σ_y) , we will use $L_I(y, \sigma_y)$ to represent the conditional likelihood of y given that its corresponding value of θ is contained in I , and $L_{\bar{I}}(y)$ for the conditional likelihood of y given that its corresponding value of θ is not contained in I . We have that the likelihood of y is $\phi L_I(y) + (1 - \phi)L_{\bar{I}}(y)$. If we change the prior to have probability α at x and $1 - \alpha$ times the previous prior, then the log-likelihood is larger than $l(\mathcal{X} \setminus (x, \sigma)) + (n-1)\log(1 - \alpha) + \log\left((1 - \alpha)L(x) + \frac{\alpha}{\sqrt{2\pi}\sigma}\right)$. The increase in log-likelihood is therefore

$$(n-1)\log(1 - \alpha) + \log\left(\frac{(1 - \alpha)L(x) + \frac{\alpha}{\sqrt{2\pi}\sigma}}{L(x)}\right) = (n-1)\log(1 - \alpha) + \log\left(1 - \alpha + \frac{\alpha}{\sqrt{2\pi}\sigma L(x)}\right)$$

For this to be an increase, we need

$$\begin{aligned} (n-1)\log(1 - \alpha) + \log\left(1 - \alpha + \frac{\alpha}{\sqrt{2\pi}\sigma L(x)}\right) &> 0 \\ (1 - n)\log(1 - \alpha) &< \log\left(1 - \alpha + \frac{\alpha}{\sqrt{2\pi}\sigma L(x)}\right) \\ (1 - \alpha)^{(1-n)} &< 1 - \alpha + \frac{\alpha}{\sqrt{2\pi}\sigma L(x)} \\ (1 - \alpha)((1 - \alpha)^{-n} - 1) &< \frac{\alpha}{\sqrt{2\pi}\sigma L(x)} \end{aligned}$$

If we substitute $\alpha = \frac{\beta}{n}$, where $\beta < 1$, then $(1 - \alpha)^{-n} \approx e^\beta$ for large n . We therefore need

$$\begin{aligned} (1 - \alpha)(e^\beta - 1) &< \frac{\alpha}{\sqrt{2\pi\sigma}L(x)} \\ \sqrt{2\pi\sigma}L(x) &< \frac{\alpha}{(1 - \alpha)(e^\beta - 1)} \\ \sqrt{2\pi\sigma}L(x) &< \frac{\beta}{(n - \beta)(e^\beta - 1)} \end{aligned}$$

Now we know that $L(x) = L_I(x) + L_{\bar{I}}(x)$, and $L_I(x) \leq \frac{\phi}{\sqrt{2\pi\sigma}}$, since the prior probability of the interval I is at most ϕ , and for each point b_i , the likelihood is $\frac{\pi_i e^{-\frac{(x-b_i)^2}{2\sigma^2}}}{(\sum_{i=1}^m \pi_i) \sqrt{2\pi\sigma}}$ so $L_I(x) = \sum_{i=1}^m \frac{\pi_i e^{-\frac{(x-b_i)^2}{2\sigma^2}}}{(\sum_{i=1}^m \pi_i) \sqrt{2\pi\sigma}} < \frac{1}{\sqrt{2\pi\sigma}}$. Also $L_{\bar{I}}(x) \leq (1 - \phi) \frac{e^{-(\log(n) + \frac{1}{2})}}{\sqrt{2\pi\sigma}} = (1 - \phi) \frac{e^{-\frac{1}{2}}}{n \sqrt{2\pi\sigma}}$ so $L(x) \leq \frac{\phi + \frac{(1-\phi)}{n} e^{-\frac{1}{2}}}{\sqrt{2\pi\sigma}}$. Therefore, provided that

$$\frac{\beta}{(n - \beta)(e^\beta - 1)} > \phi + \frac{(1 - \phi)}{n} e^{-\frac{1}{2}}$$

we will have an improvement in likelihood. we see that as $\beta \rightarrow 0$, $\frac{\beta}{e^\beta - 1} \rightarrow 1$. For small β , the left-hand side is approximately $\frac{1}{n}$. For the right-hand side, we are given that $\phi < \frac{1 - e^{-\frac{1}{2}}}{n}$, so the right-hand side is less than

$$\frac{1 - e^{-\frac{1}{2}}}{n} + \frac{e^{-\frac{1}{2}}}{n} = \frac{1}{n}$$

so the required inequality holds. □

C Optimal Parameter Values and Loss Functions for Simulations

Recall that for a distribution with density function $\pi(\theta)$, we define $\lambda(\theta) = -\frac{\pi'(\theta)}{\pi(\theta)}$, and that the best choice of estimating prior to use for ranking is chosen to minimise

$$\int_a^\infty \pi(\theta)^2 (\lambda(\theta) - \hat{\lambda}(\theta))^2 d\theta$$

where $\lambda(\theta)$ and $\pi(\theta)$ are for the true prior, while $\hat{\lambda}(\theta)$ is for the estimating prior. We evaluate this loss function for each combination of priors.

C.1 Loss functions

C.1.1 Normal Estimated by Normal

If the true prior is normal with mean 0 variance τ^2 , and the estimated prior has mean 0, variance $\hat{\tau}^2$, then the loss function is given by

$$\begin{aligned} \int_a^\infty \pi(\theta)^2 \left(\frac{\theta}{\tau^2} - \frac{\theta}{\hat{\tau}^2} \right)^2 d\theta &= \left(\frac{1}{\tau^2} - \frac{1}{\hat{\tau}^2} \right)^2 \int_a^\infty \theta^2 \frac{e^{-\frac{\theta^2}{\tau^2}}}{2\pi\tau^2} d\theta \\ &= \left(\frac{1}{\tau^2} - \frac{1}{\hat{\tau}^2} \right)^2 \left(\frac{\left[-\theta e^{-\frac{\theta^2}{\tau^2}} \right]_a^\infty}{4\pi} + \int_a^\infty \frac{e^{-\frac{\theta^2}{\tau^2}}}{4\pi} d\theta \right) \\ &= \left(\frac{1}{\tau^2} - \frac{1}{\hat{\tau}^2} \right)^2 \left(\frac{ae^{-\frac{a^2}{\tau^2}}}{4\pi} + \frac{\tau}{4\sqrt{\pi}} \left(1 - \Phi \left(\frac{\sqrt{2}a}{\tau} \right) \right) \right) \end{aligned}$$

C.1.2 Exponential estimated by normal

For the exponential true prior we have $\pi(\theta) = \lambda e^{-\lambda\theta}$ and $\lambda(\theta) = \lambda$. Meanwhile, for the normal estimating prior, we have that $\hat{\lambda}(\hat{\theta}) = \frac{\hat{\theta}}{\hat{\tau}^2}$. We are aiming to choose $\hat{\tau}$ so as to minimise

$$\begin{aligned} \int_a^\infty \pi(\theta)^2 \left(\lambda - \hat{\lambda}(\theta) \right)^2 d\theta &= \int_a^\infty \pi(\theta)^2 \left(\lambda - \frac{\theta}{\hat{\tau}^2} \right)^2 d\theta \\ &= \int_a^\infty \lambda^2 e^{-2\lambda\theta} \left(\lambda - \frac{\theta}{\hat{\tau}^2} \right)^2 d\theta \end{aligned}$$

We recall that

$$\begin{aligned} \int_a^\infty e^{-2\lambda\theta} d\theta &= \frac{e^{-2\lambda a}}{2\lambda} \\ \int_a^\infty \theta e^{-2\lambda\theta} d\theta &= \frac{e^{-2\lambda a}(2\lambda a + 1)}{4\lambda^2} \\ \int_a^\infty \theta^2 e^{-2\lambda\theta} d\theta &= \frac{e^{-2\lambda a}(2\lambda^2 a^2 + 2\lambda a + 1)}{4\lambda^3} \end{aligned}$$

Therefore, the objective function is

$$\frac{e^{-2\lambda a}}{\lambda} \left(\frac{\lambda^4}{2} - \frac{2\lambda^3 a + \lambda^2}{2\hat{\tau}^2} + \frac{2\lambda^2 a^2 + 2\lambda a + 1}{4\hat{\tau}^4} \right)$$

C.1.3 Pareto estimated by normal

For the normal estimating prior, we have $\hat{\lambda}(\theta) = \frac{\theta}{\hat{\tau}^2}$. For the Pareto true prior, we have $\lambda(\theta) = \frac{\alpha+1}{\theta}$. The objective function is therefore

$$\begin{aligned}
\int_a^\infty \pi(\theta)^2 (\lambda(\theta) - \hat{\lambda}(\theta))^2 d\theta &= \alpha^2 \int_a^\infty \frac{\eta^{2\alpha}}{\theta^{2\alpha+2}} \left(\frac{\alpha+1}{\theta} - \frac{\theta}{\hat{\tau}^2} \right)^2 d\theta \\
&= \alpha^2 \eta^{2\alpha} \int_a^\infty \left(\frac{(\alpha+1)^2}{\theta^{2\alpha+4}} - \frac{2(\alpha+1)}{\theta^{2\alpha+2}\hat{\tau}^2} + \frac{1}{\theta^{2\alpha}\hat{\tau}^4} \right) d\theta \\
&= \alpha^2 \eta^{2\alpha} \left[-\frac{(\alpha+1)^2}{(2\alpha+3)\theta^{2\alpha+3}} + \frac{2(\alpha+1)}{(2\alpha+1)\theta^{2\alpha+1}\hat{\tau}^2} - \frac{1}{(2\alpha-1)\theta^{2\alpha-1}\hat{\tau}^4} \right]_a^\infty \\
&= \frac{\alpha^2 \eta^{2\alpha}}{a^{2\alpha+3}} \left(\frac{(\alpha+1)^2}{(2\alpha+3)} - \frac{2(\alpha+1)a^2}{(2\alpha+1)\hat{\tau}^2} + \frac{a^4}{(2\alpha-1)\hat{\tau}^4} \right)
\end{aligned}$$

C.1.4 Normal estimated by exponential

The expected loss is

$$\begin{aligned}
\hat{\lambda}^2 \int_a^\infty \pi(\theta)^2 d\theta + 2\hat{\lambda} \int_a^\infty \pi(\theta)\pi'(\theta) d\theta + \int_a^\infty \pi'(\theta)^2 d\theta &= \int_a^\infty \pi'(\theta)^2 d\theta - \hat{\lambda}\pi(a)^2 + \hat{\lambda}^2 \int_a^\infty \pi(\theta)^2 d\theta \\
&= \int_a^\infty \frac{\theta^2}{2\pi\tau^6} e^{-\frac{\theta^2}{\tau^2}} d\theta - \hat{\lambda} \frac{e^{-\frac{a^2}{\tau^2}}}{2\pi\tau^2} + \hat{\lambda}^2 \int_a^\infty \frac{e^{-\frac{\theta^2}{\tau^2}}}{2\pi\tau^2} d\theta
\end{aligned}$$

We have

$$\begin{aligned}
\int_a^\infty \theta \frac{2\theta}{\tau^2} e^{-\frac{\theta^2}{\tau^2}} d\theta &= \left[-\theta e^{-\frac{\theta^2}{\tau^2}} \right]_a^\infty + \int_a^\infty e^{-\frac{\theta^2}{\tau^2}} d\theta \\
&= ae^{-\frac{a^2}{\tau^2}} + \sqrt{\pi}\tau \left(1 - \Phi\left(\frac{\sqrt{2}a}{\tau}\right) \right) \\
\int_a^\infty \frac{e^{-\frac{\theta^2}{\tau^2}}}{2\pi\tau^2} d\theta &= \frac{1 - \Phi\left(\frac{\sqrt{2}a}{\tau}\right)}{2\sqrt{\pi}\tau}
\end{aligned}$$

so the expected loss is

$$\frac{\hat{\lambda}^2}{2\sqrt{\pi}\tau} \left(1 - \Phi\left(\frac{\sqrt{2}a}{\tau}\right) \right) - \frac{\hat{\lambda}e^{-\frac{a^2}{\tau^2}}}{2\pi\tau^2} + \frac{ae^{-\frac{a^2}{\tau^2}}}{4\pi\tau^4} + \frac{1 - \Phi\left(\frac{\sqrt{2}a}{\tau}\right)}{4\sqrt{\pi}\tau^3}$$

C.1.5 Exponential Estimated by Exponential

If the true prior is exponential with rate λ , and the estimated prior is exponential with rate $\hat{\lambda}$, then the loss function is given by

$$\begin{aligned}\int_a^\infty \pi(\theta)^2 (\lambda - \hat{\lambda})^2 d\theta &= (\lambda - \hat{\lambda})^2 \int_a^\infty \lambda^2 e^{-2\lambda\theta} d\theta \\ &= (\lambda - \hat{\lambda})^2 \left[\frac{-\lambda e^{-2\lambda\theta}}{2} \right]_a^\infty \\ &= \frac{\lambda (\lambda - \hat{\lambda})^2}{2} e^{-2\lambda a}\end{aligned}$$

C.1.6 Pareto estimated by exponential

The loss function is

$$\begin{aligned}\hat{\lambda}^2 \int_a^\infty \left(\frac{\alpha \eta^\alpha}{\theta^{\alpha+1}} \right)^2 d\theta - 2\hat{\lambda} \int_a^\infty \left(\frac{\alpha \eta^\alpha}{\theta^{\alpha+1}} \right) \left(\frac{\alpha(\alpha+1)\eta^\alpha}{\theta^{\alpha+2}} \right) d\theta + \int_a^\infty \left(\frac{\alpha(\alpha+1)\eta^\alpha}{\theta^{\alpha+2}} \right)^2 d\theta &= \frac{\hat{\lambda}^2 \alpha^2 \eta^{2\alpha}}{(2\alpha+1)a^{2\alpha+1}} - 2 \frac{\hat{\lambda} \alpha^2 (\alpha+1) \eta^{2\alpha}}{(2\alpha+2)a^{2\alpha+2}} + \frac{\alpha^2 (\alpha+1)^2 \eta^{2\alpha}}{(2\alpha+3)a^{2\alpha+3}} \\ &= \frac{\alpha^2 \eta^{2\alpha}}{a^{2\alpha+1}} \left(\frac{\hat{\lambda}^2}{(2\alpha+1)} - \frac{\hat{\lambda}}{a} + \frac{(\alpha+1)^2}{(2\alpha+3)a^2} \right)\end{aligned}$$

C.1.7 Normal Estimated by Pareto

For the Normal estimated by Pareto, we have $\hat{\lambda}(\theta) = \frac{\hat{\alpha}+1}{\theta}$. The loss function is therefore

$$\begin{aligned}\frac{1}{2\pi\tau^2} \int_a^\infty e^{-\frac{\theta^2}{\tau^2}} \left(\frac{\theta}{\tau^2} - \frac{\hat{\alpha}+1}{\theta} \right)^2 d\theta &= \frac{1}{2\pi\tau^2} \int_a^\infty e^{-\frac{\theta^2}{\tau^2}} \left(\frac{\theta^2}{\tau^4} - \frac{2(\hat{\alpha}+1)}{\tau^2} + \frac{(\hat{\alpha}+1)^2}{\theta^2} \right) d\theta \\ &= \frac{1}{2\pi\tau^2} \left((\hat{\alpha}+1)^2 \int_a^\infty \frac{1}{\theta^2} e^{-\frac{\theta^2}{\tau^2}} d\theta - 2(\hat{\alpha}+1) \int_a^\infty \frac{1}{\tau^2} e^{-\frac{\theta^2}{\tau^2}} d\theta + \int_a^\infty \frac{\theta^2}{\tau^4} e^{-\frac{\theta^2}{\tau^2}} d\theta \right) \\ &= \frac{1}{2\pi\tau^2} \left((\hat{\alpha}+1)^2 \int_a^\infty \frac{1}{\theta^2} e^{-\frac{\theta^2}{\tau^2}} d\theta - 2(\hat{\alpha}+1) \int_a^\infty \frac{1}{\tau^2} e^{-\frac{\theta^2}{\tau^2}} d\theta + \int_a^\infty \frac{\theta}{\tau^4} \theta e^{-\frac{\theta^2}{\tau^2}} d\theta \right) \\ &= \frac{1}{2\pi\tau^2} \left((\hat{\alpha}+1)^2 \int_a^\infty \frac{1}{\theta^2} e^{-\frac{\theta^2}{\tau^2}} d\theta - 2(\hat{\alpha}+1) \int_a^\infty \frac{1}{\tau^2} e^{-\frac{\theta^2}{\tau^2}} d\theta + \left[-\frac{\theta}{2\tau^2} e^{-\frac{\theta^2}{\tau^2}} \right]_a^\infty + \int_a^\infty \frac{1}{2\tau^2} e^{-\frac{\theta^2}{\tau^2}} d\theta \right) \\ &= \frac{1}{2\pi\tau^2} \left((\hat{\alpha}+1)^2 \int_a^\infty \frac{1}{\theta^2} e^{-\frac{\theta^2}{\tau^2}} d\theta - \left(2\hat{\alpha} + \frac{3}{2} \right) \frac{\sqrt{\pi}}{\tau} \left(1 - \Phi \left(\frac{\sqrt{2}a}{\tau} \right) \right) + \frac{a}{2\tau^2} e^{-\frac{a^2}{\tau^2}} \right)\end{aligned}$$

C.1.8 Exponential estimated by Pareto

For the Exponential estimated by Pareto, we have $\hat{\lambda}(\theta) = \frac{\hat{\alpha}+1}{\theta}$. The loss is therefore

$$\begin{aligned} \lambda^2 \int_a^\infty e^{-2\lambda\theta} \left(\lambda - \frac{\hat{\alpha}+1}{\theta} \right)^2 d\theta &= \lambda^2 \int_a^\infty e^{-2\lambda\theta} \left(\lambda^2 - \frac{2\lambda(\hat{\alpha}+1)}{\theta} + \frac{(\hat{\alpha}+1)^2}{\theta^2} \right) d\theta \\ &= \lambda^2 \left((\hat{\alpha}+1)^2 \int_a^\infty \frac{1}{\theta^2} e^{-2\lambda\theta} d\theta - 2\lambda(\hat{\alpha}+1) \int_a^\infty \frac{1}{\theta} e^{-2\lambda\theta} d\theta + \lambda^2 \int_a^\infty e^{-2\lambda\theta} d\theta \right) \end{aligned}$$

C.1.9 Pareto Estimated by Pareto

If the true prior is Pareto with minimum η and index α , and the estimated prior is Pareto with minimum η and index $\hat{\alpha}$, then the loss function is given by

$$\begin{aligned} \int_a^\infty \pi(\theta)^2 \left(\frac{\alpha+1}{\theta} - \frac{\hat{\alpha}+1}{\theta} \right)^2 d\theta &= (\alpha - \hat{\alpha})^2 \int_a^\infty \frac{\alpha^2 \eta^{2\alpha}}{\theta^{2\alpha+4}} d\theta \\ &= (\alpha - \hat{\alpha})^2 \left[-\frac{\alpha^2 \eta^{2\alpha}}{(2\alpha+3)\theta^{2\alpha+3}} \right]_a^\infty \\ &= (\alpha - \hat{\alpha})^2 \frac{\alpha^2 \eta^{2\alpha}}{(2\alpha+3)a^{2\alpha+3}} \end{aligned}$$

C.1.10 MLE Ranking with Normal Prior

For a normal true prior, we have that the loss from using the MLE ranking is

$$\int_a^\infty \pi'(\theta)^2 d\theta = \int_a^\infty \frac{\theta^2}{2\pi\tau^6} e^{-\frac{\theta^2}{\tau^2}} d\theta = \frac{1}{4\pi\tau^4} \left(\left[-\theta e^{-\frac{\theta^2}{\tau^2}} \right]_a^\infty + \int_a^\infty e^{-\frac{\theta^2}{\tau^2}} d\theta \right) = \frac{1}{4\pi\tau^4} \left(a e^{-\frac{a^2}{\tau^2}} + \sqrt{\pi}\tau \left(1 - \Phi \left(\frac{\sqrt{2}a}{\tau} \right) \right) \right)$$

For $\tau = 1$, $a = 1.281552$ this loss is 0.02466714.

C.1.11 MLE Ranking with Exponential Prior

The expected loss function using the MLE ranking is

$$\int_a^\infty \pi'(\theta)^2 d\theta = \int_a^\infty (-\lambda^2 e^{-\lambda\theta})^2 d\theta = \lambda^4 \int_a^\infty e^{-2\lambda\theta} d\theta = \frac{\lambda^3}{2} \left[-e^{-2\lambda\theta} \right]_a^\infty = \frac{\lambda^3 e^{-2\lambda a}}{2}$$

Substituting $\lambda = 1$ and $a = \log(10)$ this gives 0.005.

C.1.12 MLE Ranking with Pareto Prior

For the Pareto true prior, the expected additional loss from using the MLE ranking is

$$\int_a^\infty \pi'(\theta)^2 d\theta = \int_a^\infty \left(\frac{\alpha(\alpha+1)\eta^\alpha}{\theta^{\alpha+2}} \right)^2 d\theta = \alpha^2(\alpha+1)^2\eta^{2\alpha} \int_a^\infty \theta^{-(2\alpha+4)} d\theta = \frac{\alpha^2(\alpha+1)^2\eta^{2\alpha}}{2\alpha+3} \left[-\theta^{-(2\alpha+3)} \right]_a^\infty = \frac{\alpha^2(\alpha+1)^2\eta^{2\alpha}}{(2\alpha+3)a^{2\alpha+3}}$$

Substituting $\alpha = 2$, $\eta = \frac{1}{2}$ and $a = \frac{\sqrt{10}}{2}$, we get the loss is

$$\frac{2^2 3^2 \left(\frac{1}{2}\right)^4}{7 \left(\frac{\sqrt{10}}{2}\right)^7} = \frac{288}{7000 \sqrt{10}} = 0.01301051$$

C.2 Optimal Parameter estimates

C.2.1 Exponential estimated by normal

The loss function is minimised by

$$\begin{aligned} \frac{1}{\tau^2} &= \frac{2\lambda^3 a + \lambda^2}{2\lambda^2 a^2 + 2\lambda a + 1} \\ \tau^2 &= \frac{1}{\lambda^2} \left(\frac{2\lambda^2 a^2 + 2\lambda a + 1}{2\lambda a + 1} \right) \end{aligned}$$

Substituting $\lambda = 1$ and $a = \log(10)$ (the 90th percentile of the exponential distribution) gives

$$\hat{\tau} = \sqrt{\frac{2 \log(10)^2 + 2 \log(10) + 1}{2 \log(10) + 1}} = 1.700526$$

and the expected loss is

$$0.01 \left(\frac{1}{2} - \frac{(2 \log(10) + 1)^2}{4(2 \log(10)^2 + 2 \log(10) + 1)} \right) = 0.0001542356$$

C.2.2 Pareto estimated by normal

The loss function is minimised by

$$\begin{aligned} \frac{1}{\tau^2} &= \frac{\left(\frac{\alpha+1}{(2\alpha+1)a^{2\alpha+1}} \right)}{\left(\frac{1}{(2\alpha-1)a^{2\alpha-1}} \right)} = \frac{(2\alpha-1)(\alpha+1)}{(2\alpha+1)a^2} \\ \tau^2 &= \frac{2\alpha+1}{(\alpha+1)(2\alpha-1)} a^2 \end{aligned}$$

For this value, the loss is

$$\frac{4\alpha^2(\alpha+1)^2\eta^{2\alpha}}{(2\alpha+3)(2\alpha+1)^2a^{2\alpha+3}}$$

Substituting the values $\alpha = 2$, $\eta = \frac{1}{2}$ used in the simulation and the corresponding 90th percentile $a = \frac{\sqrt{10}}{2}$, we get that the optimal parameter τ has

$$\frac{1}{\tau^2} = \frac{(2\alpha-1)(\alpha+1)}{(2\alpha+1)a^2} = \frac{3 \times 3}{5 \times \frac{10}{4}} = 0.72$$

and the loss is

$$\frac{4\alpha^2(\alpha+1)^2\eta^{2\alpha}}{(2\alpha+3)(2\alpha+1)^2a^{2\alpha+3}} = \frac{4 \times 2^2 \times 3^2 \times \left(\frac{1}{2}\right)^4}{7 \times 5^2 \times \left(\frac{\sqrt{10}}{2}\right)^7} = 0.002081682$$

C.2.3 Normal estimated by exponential

If the true prior is normal, but we are using an exponential, then recall that the best choice is

$$\lambda = \frac{\pi(a)^2}{2 \int_a^\infty \pi(\theta)^2 d\theta}$$

We evaluate

$$\int_a^\infty \pi(\theta)^2 d\theta = \frac{1}{2\pi\tau^2} \int_a^\infty e^{-\frac{\theta^2}{\tau^2}} d\theta = \frac{1}{2\sqrt{\pi}\tau} \int_a^\infty \frac{1}{\sqrt{2\pi}\frac{\tau}{\sqrt{2}}} e^{-\frac{\theta^2}{\tau^2}} d\theta = \frac{1}{2\sqrt{\pi}\tau} \left(1 - \Phi\left(\frac{\sqrt{2}a}{\tau}\right)\right)$$

so the best choice of $\hat{\lambda}$ for the exponential estimating prior is

$$\hat{\lambda} = \frac{e^{-\frac{a^2}{\tau^2}}}{2\sqrt{\pi}\tau \left(1 - \Phi\left(\frac{\sqrt{2}a}{\tau}\right)\right)}$$

For the simulation setting $\tau = 1$, $a = 1.281552$, this is $\hat{\lambda} = 1.561386$ and the expected loss for our simulation is 0.000622064.

C.2.4 Pareto estimated by exponential

For the exponential prior, the best choice of λ is given by

$$\lambda = \frac{\pi(a)^2}{2 \int_a^\infty \pi(\theta)^2 d\theta}$$

We evaluate

$$\int_a^\infty \pi(\theta)^2 d\theta = \alpha^2 \int_a^\infty \frac{\eta^{2\alpha}}{\theta^{2\alpha+2}} d\theta = \alpha^2 \left[\frac{-\eta^{2\alpha}}{(2\alpha+1)\theta^{2\alpha+1}} \right]_a^\infty = \alpha^2 \frac{\eta^{2\alpha}}{(2\alpha+1)a^{2\alpha+1}}$$

so the best choice of $\hat{\lambda}$ for the exponential estimating prior is

$$\hat{\lambda} = \frac{\alpha^2 \left(\frac{\eta^{2\alpha}}{a^{2\alpha+2}} \right)}{2\alpha^2 \left(\frac{\eta^{2\alpha}}{(2\alpha+1)a^{2\alpha+1}} \right)} = \frac{2\alpha+1}{2a}$$

for this $\hat{\lambda}$ the expected loss is

$$\frac{\alpha^2 \eta^{2\alpha}}{4(2\alpha+3)a^{2\alpha+3}}$$

Substituting the values $\alpha = 2$, $\eta = \frac{1}{2}$ and the corresponding 90th percentile $a = \frac{\sqrt{10}}{2}$, we get $\hat{\lambda} = 1.581139$ and the loss is

$$\frac{2^2 \left(\frac{1}{2} \right)^4}{4 \times 7 \left(\frac{\sqrt{10}}{2} \right)^7} = \frac{8}{7000 \sqrt{10}} = 0.0003614032$$

C.2.5 Normal Estimated by Pareto

For the Normal estimated by Pareto, the loss is minimised by

$$\alpha + 1 = \frac{\int_a^\infty \frac{1}{\tau^2} e^{-\frac{\theta^2}{\tau^2}} d\theta}{\int_a^\infty \frac{1}{\theta^2} e^{-\frac{\theta^2}{\tau^2}} d\theta}$$

For this choice of α , the loss is

$$\frac{1}{2\pi\tau^2} \left(\int_a^\infty \frac{\theta^2}{\tau^4} e^{-\frac{\theta^2}{\tau^2}} d\theta - \frac{\left(\int_a^\infty \frac{1}{\tau^2} e^{-\frac{\theta^2}{\tau^2}} d\theta \right)^2}{\int_a^\infty \frac{1}{\theta^2} e^{-\frac{\theta^2}{\tau^2}} d\theta} \right)$$

For the case in our simulation, we have $\tau = 1$ and $a = 1.281552$. For these values we calculate numerically

$$\begin{aligned}\int_a^\infty \frac{1}{\theta^2} e^{-\frac{\theta^2}{\tau^2}} d\theta &= 0.02706327 \\ \int_a^\infty \frac{\theta^2}{\tau^4} e^{-\frac{\theta^2}{\tau^2}} d\theta &= 0.1549882 \\ \int_a^\infty \frac{1}{\tau^2} e^{-\frac{\theta^2}{\tau^2}} d\theta &= \sqrt{\pi}(1 - \Phi(\sqrt{2}a)) = 0.06197059\end{aligned}$$

Substituting these into the formula, we get that the expected loss is

$$\frac{1}{2\pi} \left(0.1549882 - \frac{(0.06197059)^2}{0.02706327} \right) = 0.002082605$$

C.2.6 Exponential estimated by Pareto

The loss is minimised by

$$\alpha + 1 = \frac{\lambda \int_a^\infty \frac{1}{\theta} e^{-2\lambda\theta} d\theta}{\int_a^\infty \frac{1}{\theta^2} e^{-2\lambda\theta} d\theta}$$

Integrating by parts gives

$$\int_a^\infty \frac{1}{\theta} e^{-2\lambda\theta} d\theta = \left[-\frac{1}{2\lambda\theta} e^{-2\lambda\theta} \right]_a^\infty - \int_a^\infty \frac{1}{2\lambda\theta^2} e^{-2\lambda\theta} d\theta = \frac{e^{-2\lambda a}}{2\lambda a} - \frac{1}{2\lambda} \int_a^\infty \frac{1}{\theta^2} e^{-2\lambda\theta} d\theta$$

We therefore get

$$\begin{aligned}\alpha + 1 &= \frac{\frac{e^{-2\lambda a}}{2a} - \frac{1}{2} \int_a^\infty \frac{1}{\theta^2} e^{-2\lambda\theta} d\theta}{\int_a^\infty \frac{1}{\theta^2} e^{-2\lambda\theta} d\theta} = \frac{e^{-2\lambda a}}{2a \int_a^\infty \frac{1}{\theta^2} e^{-2\lambda\theta} d\theta} - \frac{1}{2} \\ \alpha &= \frac{e^{-2\lambda a}}{2a \int_a^\infty \frac{1}{\theta^2} e^{-2\lambda\theta} d\theta} - \frac{3}{2}\end{aligned}$$

We have $\lambda = 1$ and $a = \log(10)$, so numerically we obtain

$$\begin{aligned}\int_a^\infty \frac{1}{\theta} e^{-2\lambda\theta} d\theta &= 0.001829743 \\ \int_a^\infty \frac{1}{\theta^2} e^{-2\lambda\theta} d\theta &= 0.0006834578 \\ \int_a^\infty e^{-2\lambda\theta} d\theta &= \frac{e^{-2\lambda a}}{2\lambda} = 0.005\end{aligned}$$

This gives the optimal parameter estimate as

$$\hat{\alpha} = \frac{0.01}{2 \times 0.0006834578 \log(10)} - \frac{3}{2} = 1.677186$$

so the expected loss is 0.0001014394

C.3 MLE Estimates for Parameters of Estimating Priors

We will assume that a is given for each simulation, and that our objective is to estimate the parameters from the data for each estimating prior so that the distribution fits the data well on the tail. We will use maximum likelihood for this purpose. We have already seen that the loss function is different from the Kullback-Leibler divergence that the MLE estimate attempts to optimise, so the MLE is not optimal in terms of minimising our expected loss function, and further work could go into devising better estimation methods for the misspecified prior case. For the MLE estimation, the details in each case are presented here:

C.3.1 Normal Distribution

We have n samples which we model as having mean θ_i following a normal distribution with mean 0 and variance τ^2 , and each observation x_i following a normal distribution with mean θ_i and variance σ_i^2 . We want to maximise the log-likelihood of all the data points with $\theta_i > a$ for some cutoff a . To simplify this procedure, we will maximise the log-likelihood of all data points for which $x_i > a$. The log-likelihood is then written

$$\sum_{x_i > a} \left(-\frac{x_i^2}{2(\tau^2 + \sigma_i^2)} - \frac{\log(\tau^2 + \sigma_i^2)}{2} - \log \left(1 - \Phi \left(\frac{a}{\sqrt{\tau^2 + \sigma_i^2}} \right) \right) \right)$$

(The last term is because we must take the conditional log-likelihood conditional on $x_i > a$.) Setting the derivative with respect to τ to zero, we get

$$\sum_{x_i > a} \left(\frac{\tau x_i^2}{(\tau^2 + \sigma_i^2)^2} - \frac{\tau}{(\tau^2 + \sigma_i^2)} - \frac{\tau a e^{-\frac{a^2}{2(\tau^2 + \sigma_i^2)}}}{\sqrt{2\pi}(\tau^2 + \sigma_i^2)^{\frac{3}{2}} \left(1 - \Phi \left(\frac{a}{\sqrt{\tau^2 + \sigma_i^2}} \right) \right)} \right) = 0$$

We can solve this numerically using Newton's method. We can use the following method to obtain a good starting value. Since a is reasonably large compared to τ , we can approximate

$$\frac{e^{-\frac{a^2}{2(\tau^2 + \sigma_i^2)}}}{\sqrt{2\pi}(\tau^2 + \sigma_i^2)^{\frac{3}{2}} \left(1 - \Phi \left(\frac{a}{\sqrt{\tau^2 + \sigma_i^2}} \right) \right)} \approx \frac{a}{\tau^2 + \sigma_i^2}$$

[NOTE: this is a poor approximation. Using it gives fairly bad approximations for $\hat{\tau}$. The approximations for other estimating priors later are better.] so that the final term in the derivative of the log-likelihood is approximately

$$\frac{\tau a^2}{(\tau^2 + \sigma_i^2)^2}$$

We have assumed that σ_i is small compared to τ , so we can set

$$\begin{aligned}
\sum_{x_i > a} \left(\frac{\tau(x_i^2 - a^2)}{(\tau^2 + \sigma_i^2)^2} - \frac{\tau}{(\tau^2 + \sigma_i^2)} \right) &= \sum_{x_i > a} \left(\tau^{-3}(x_i^2 - a^2) \left(1 + \frac{\sigma_i^2}{\tau^2} \right)^{-2} - \tau^{-1} \left(1 + \frac{\sigma_i^2}{\tau^2} \right)^{-1} \right) \\
&\approx \sum_{x_i > a} \left(\tau^{-3}(x_i^2 - a^2) \left(1 - 2\frac{\sigma_i^2}{\tau^2} \right) - \tau^{-1} \left(1 - \frac{\sigma_i^2}{\tau^2} \right) \right) \\
&= \tau^{-5} \sum_{x_i > a} \left(-\tau^4 + \tau^2((x_i^2 - a^2) + \sigma_i^2) - 2\sigma_i^2(x_i^2 - a^2) \right) \\
&= \tau^{-5} \left(-n_a \tau^4 + \tau^2 \sum_{x_i > a} ((x_i^2 - a^2) + \sigma_i^2) - 2 \sum_{x_i > a} \sigma_i^2(x_i^2 - a^2) \right)
\end{aligned}$$

where n_a is the number of points with $x_i > a$.

We solve for when this is equal to zero using the quadratic formula to get:

$$\tau^2 \approx \frac{\sum_{x_i > a} ((x_i^2 - a^2) + \sigma_i^2) + \sqrt{\left(\sum_{x_i > a} ((x_i^2 - a^2) + \sigma_i^2) \right)^2 - 8n_a \sum_{x_i > a} (x_i^2 - a^2) \sigma_i^2}}{2n_a}$$

which should give an approximation to the true value of τ . If we further make the approximation that $\frac{8n_a \sum_{x_i > a} (x_i^2 - a^2) \sigma_i^2}{\left(\sum_{x_i > a} ((x_i^2 - a^2) + \sigma_i^2) \right)^2}$ is small, then we have

$$\sqrt{\left(\sum_{x_i > a} ((x_i^2 - a^2) + \sigma_i^2) \right)^2 - 8n_a \sum_{x_i > a} (x_i^2 - a^2) \sigma_i^2} \approx \sum_{x_i > a} ((x_i^2 - a^2) + \sigma_i^2) - \frac{8n_a \sum_{x_i > a} (x_i^2 - a^2) \sigma_i^2}{2 \sum_{x_i > a} ((x_i^2 - a^2) + \sigma_i^2)}$$

which gives us

$$\tau^2 \approx \frac{\sum_{x_i > a} ((x_i^2 - a^2) + \sigma_i^2)}{n_a} - \frac{2 \sum_{x_i > a} (x_i^2 - a^2) \sigma_i^2}{\sum_{x_i > a} ((x_i^2 - a^2) + \sigma_i^2)}$$

We can compare this approximate MLE estimate of τ^2 to the theoretically best estimate for the exponential and Pareto cases. If we assume that σ_i are all small, then the term $\frac{\sum_{x_i > a} (x_i^2 - a^2) \sigma_i^2}{\sum_{x_i > a} ((x_i^2 - a^2) + \sigma_i^2)}$ is approximately $\sum_{x_i > a} \frac{(x_i^2 - a^2)}{\sum_{x_i > a} (x_i^2 - a^2)} \sigma_i^2$, which is a weighted mean of the σ_i^2 . Therefore the expected value is the expected value of σ_i^2 , so we have

$$\mathbb{E}(\hat{\tau}^2) \approx \mathbb{E}((x_i^2 - a^2)) - \mathbb{E}(\sigma_i^2)$$

Since x_i is normally distributed with mean θ_i and variance σ_i^2 , we have that

$$\mathbb{E}(x_i^2 | \theta_i) = (\mathbb{E}(x_i | \theta_i))^2 + \sigma_i^2 = \theta_i^2 + \sigma_i^2$$

Therefore we have

$$\mathbb{E}(\hat{\tau}^2) \approx \mathbb{E}_{x_i > a}(\theta_i^2 - a^2) \approx \mathbb{E}_{\theta_i > a}(\theta_i^2 - a^2)$$

For the exponential true prior, we have that conditional on $\theta_i > a$, we have $T = \theta_i - a$ follows an exponential distribution with $\lambda = 1$ and $a = \log(10)$. Therefore

$$\begin{aligned}
\mathbb{E}_{\theta_i > a}(\theta_i^2) &= \mathbb{E}((T + a)^2) = a^2 + 2a\mathbb{E}(T) + \mathbb{E}(T^2) = a^2 + 2\frac{a}{\lambda} + \frac{2}{\lambda^2} \\
\mathbb{E}_{\theta_i > a}(\theta_i^2) - a^2 &= 2\frac{a}{\lambda} + \frac{2}{\lambda^2} = 2\log(10) + 2 = 6.60517
\end{aligned}$$

Therefore, for a large sample

$$\hat{\tau} \approx \sqrt{6.60517} = 2.570053$$

This is quite far from the optimal estimate of 1.700526.

For the Pareto true prior, the variance is infinite (since $\alpha \leq 2$), so the distribution of the MLE $\hat{\tau}^2$ has infinite mean. This means we cannot apply the law of large numbers to assert that for large sample size $\hat{\tau}^2$ will converge in distribution to a constant. More specifically, θ_i^2 follows a Pareto distribution with $\alpha = 1$ and $\eta = \frac{1}{4}$. The sum of Pareto distributions with small α is approximately equal to the maximum value, which has distribution function

$$F_{\sum \theta_i^2}(x) = \left(1 - \frac{\eta}{x}\right)^n$$

We also have

$$F_{\hat{\tau}^2}(x) = F_{n_a \hat{\tau}^2}(n_a x) = F_{\sum \theta_i^2}(n_a(x + a^2)) = \left(1 - \frac{\eta}{n_a(x + a^2)}\right)^n \approx e^{-\frac{\eta n}{(x+a^2)n_a}}$$

We are interested in $\frac{1}{\hat{\tau}^2}$, because this is the value that is important for our posterior mean estimate. The survival function of $\frac{1}{\hat{\tau}^2}$ is

$$S_{\frac{1}{\hat{\tau}^2}}(x) = F_{\hat{\tau}^2}\left(\frac{1}{x}\right) \approx e^{-\frac{\eta n x}{n_a(1+a^2 x)}}$$

That is, $\frac{1}{\hat{\tau}^2}$ approximately follows an exponential distribution with parameter $\frac{1}{4P(\theta_i > a)} = 2.5$. This can be quite different from the optimal 0.72. Indeed we get

$$\mathbb{E}\left(\left(\frac{1}{\hat{\tau}^2} - 0.72\right)^2\right) = \mathbb{E}\left(\left(\frac{1}{\hat{\tau}^2} - 0.4\right)^2\right) + (0.4 - 0.72)^2 = \text{Var}\left(\frac{1}{\hat{\tau}^2}\right) + 0.32^2 = 0.16 + 0.1024 = 0.2624$$

Meaning that the MLE estimate for τ^2 does not give a good estimate.

C.3.2 Exponential Distribution

The likelihood of a point (x_i, σ_i) is

$$\begin{aligned} \int_0^\infty \lambda e^{-\lambda \theta} \frac{e^{-\frac{(x_i - \theta)^2}{2\sigma_i^2}}}{\sqrt{2\pi}\sigma_i} d\theta &= \frac{\lambda e^{\frac{\lambda^2 \sigma_i^2}{2} - \lambda x_i}}{\sqrt{2\pi}\sigma_i} \int_0^\infty e^{-\frac{(\theta + \lambda \sigma_i^2 - x_i)^2}{2\sigma_i^2}} d\theta \\ &= \lambda e^{\frac{\lambda^2 \sigma_i^2}{2} - \lambda x_i} \Phi\left(\frac{x_i}{\sigma_i} - \lambda \sigma_i\right) \end{aligned}$$

Since $x_i > a$ and σ_i is small, we can approximate

$$\Phi\left(\frac{x_i}{\sigma_i} - \lambda \sigma_i\right) \approx 1$$

so the log-likelihood is approximately

$$\sum_{x_i > a} \left(\log(\lambda) + \frac{\lambda^2 \sigma_i^2}{2} - \lambda x_i \right) = \lambda^2 \sum_{x_i > a} \frac{\sigma_i^2}{2} - \lambda \sum_{x_i > a} x_i + n_a \log(\lambda)$$

However, we want the conditional log-likelihood given $x_i > a$. Since σ_i is small, we will set this approximately equal to the likelihood conditional on $\theta_i > a$, which is

$$\sum_{x_i > a} \left(\log(\lambda) + \frac{\lambda^2 \sigma_i^2}{2} - \lambda x_i \right) = \lambda^2 \sum_{x_i > a} \frac{\sigma_i^2}{2} - \lambda \sum_{x_i > a} (x_i - a) + n_a \log(\lambda)$$

Setting the derivative with respect to λ to zero gives

$$\begin{aligned} \lambda \sum_{x_i > a} \sigma_i^2 - \sum_{x_i > a} (x_i - a) + \frac{n_a}{\lambda} &= 0 \\ \lambda^2 \sum_{x_i > a} \sigma_i^2 - \lambda \sum_{x_i > a} (x_i - a) + n_a &= 0 \\ \lambda &= \frac{\sum_{x_i > a} (x_i - a) \pm \sqrt{\left(\sum_{x_i > a} (x_i - a) \right)^2 - 4n_a \sum_{x_i > a} \sigma_i^2}}{2 \sum_{x_i > a} \sigma_i^2} \end{aligned}$$

so the log-likelihood is maximised by

$$\lambda = \frac{\sum_{x_i > a} (x_i - a) - \sqrt{\left(\sum_{x_i > a} (x_i - a) \right)^2 - 4n_a \sum_{x_i > a} \sigma_i^2}}{2 \sum_{x_i > a} \sigma_i^2}$$

(the other zero is because the approximation

$$\Phi\left(\frac{x_i}{\sigma_i} - \lambda \sigma_i\right) \approx 1$$

does not hold for $\lambda \approx \frac{\sum_{x_i > a} x_i}{\sum_{x_i > a} \sigma_i^2}$) Since σ_i is small, we can approximate

$$\sqrt{\left(\sum_{x_i > a} (x_i - a) \right)^2 - 4n_a \sum_{x_i > a} \sigma_i^2} \approx \sum_{x_i > a} (x_i - a) - \frac{4n_a \sum_{x_i > a} \sigma_i^2}{2 \sum_{x_i > a} (x_i - a)}$$

Which gives

$$\hat{\lambda} \approx \frac{4n_a \sum_{x_i > a} \sigma_i^2}{4 \left(\sum_{x_i > a} (x_i - a) \right) \left(\sum_{x_i > a} \sigma_i^2 \right)} = \frac{n_a}{\sum_{x_i > a} (x_i - a)}$$

When the true prior is normal, we see that $\mathbb{E}(x_i - a | x_i > a)$ is the mean of a truncated normal distribution, and is given by

$$\tau \frac{e^{-\frac{a^2}{2\tau^2}}}{\sqrt{2\pi} \left(1 - \Phi\left(\frac{a}{\tau}\right) \right)} - a$$

Substituting $\tau = 1$ and $\Phi(a) = 0.9$, we get that $\mathbb{E}(x_i - a | x_i > a) = \frac{e^{-\frac{1.281552^2}{2}}}{0.1 \sqrt{2\pi}} - a = 1.754982 - 1.281552 = 0.4734308$

Therefore, for a large sample, our estimate $\hat{\lambda}$ will converge to $\frac{1}{0.4734308} = 2.112241$.

For the Pareto true prior, we have $\mathbb{E}(x_i | x_i > a) = 2a$. Despite the fact that the variance is infinite, the law of large numbers still ensures that the sample mean of the x_i does converge to $2a$ as sample size tends to infinity. We can therefore substitute $2a$ for this sum in the expression to get

$$\hat{\lambda} \approx \frac{n_a}{\sum_{x_i > a} (x_i - a)} = \frac{1}{a} = \frac{2}{\sqrt{10}} = 0.6324555$$

C.3.3 Pareto Distribution

For the Pareto estimating prior, the likelihood of θ_i is

$$\alpha \frac{\eta^\alpha}{\theta_i^{\alpha+1}}$$

and the probability of a value exceeding a is $\frac{\eta^\alpha}{a^\alpha}$. The likelihood of (x_i, σ_i) is therefore

$$\int_{\eta}^{\infty} \alpha \frac{\eta^\alpha}{\theta^{\alpha+1}} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-\theta)^2}{2\sigma_i^2}} d\theta$$

Letting $\frac{\theta}{x_i} = 1 + \xi$, this integral becomes

$$\begin{aligned} \alpha \frac{\eta^\alpha}{x_i^{\alpha+1}} \frac{1}{\sqrt{2\pi}\sigma_i} \int_{\eta}^{\infty} (1 + \xi)^{-\alpha-1} e^{-\frac{x_i^2 \xi^2}{2\sigma_i^2}} d\theta &= \alpha \frac{\eta^\alpha}{x_i^{\alpha+1}} \frac{1}{\sqrt{2\pi}\sigma_i} \int_{\frac{\eta}{x_i}-1}^{\infty} (1 + \xi)^{-\alpha-1} e^{-\frac{x_i^2 \xi^2}{2\sigma_i^2}} d\xi \\ &= \alpha \frac{\eta^\alpha}{x_i^{\alpha+1}} \int_{\frac{\eta}{x_i}-1}^{\infty} \left(1 - (\alpha+1)\xi + \frac{(\alpha+1)(\alpha+2)}{2}\xi^2 - \dots \right) \frac{x_i e^{-\frac{x_i^2 \xi^2}{2\sigma_i^2}}}{\sqrt{2\pi}\sigma_i} d\xi \\ &\approx \alpha \frac{\eta^\alpha}{x_i^{\alpha+1}} \mathbb{E}_{\xi \sim N\left(0, \frac{\sigma_i^2}{x_i^2}\right)} \left(1 - (\alpha+1)\xi + \frac{(\alpha+1)(\alpha+2)}{2}\xi^2 - \dots \right) \\ &\approx \alpha \frac{\eta^\alpha}{x_i^{\alpha+1}} \left(1 + \frac{(\alpha+1)(\alpha+2)\sigma_i^2}{2x_i^2} + \dots \right) \end{aligned}$$

so the conditional likelihood of x_i given that $\theta_i > a$ is approximately

$$\alpha \frac{a^\alpha}{x_i^{\alpha+1}} \left(1 + \frac{(\alpha+1)(\alpha+2)\sigma_i^2}{2x_i^2} \right)$$

The conditional log-likelihood is therefore

$$\alpha \log(a) - (\alpha+1) \log(x_i) + \log(\alpha) + \log \left(1 + \frac{(\alpha+1)(\alpha+2)\sigma_i^2}{2x_i^2} \right)$$

Setting the derivative with respect to α to zero gives

$$\begin{aligned} \sum_{x_i > a} \left(\log(a) - \log(x_i) + \frac{1}{\alpha} + \frac{(2\alpha+3)\sigma_i^2}{2x_i^2 \left(1 + \frac{(\alpha+1)(\alpha+2)\sigma_i^2}{2x_i^2} \right)} \right) &= 0 \\ \sum_{x_i > a} \left(\log(a) - \log(x_i) + \frac{1}{\alpha} + \frac{(2\alpha+3)\sigma_i^2}{2x_i^2 + (\alpha+1)(\alpha+2)\sigma_i^2} \right) &= 0 \\ \frac{n_a}{\alpha} + \sum_{x_i > a} (\log(a) - \log(x_i)) + (2\alpha+3) \sum_{x_i > a} \frac{\sigma_i^2}{2x_i^2} - (\alpha+1)(\alpha+2)(2\alpha+3) \sum_{x_i > a} \frac{\sigma_i^4}{4x_i^4} &= 0 \\ n_a + \alpha \sum_{x_i > a} \left(\log(a) - \log(x_i) + 3 \frac{\sigma_i^2}{2x_i^2} - 6 \frac{\sigma_i^4}{4x_i^4} \right) + \alpha^2 \sum_{x_i > a} \left(\frac{\sigma_i^2}{x_i^2} - 13 \frac{\sigma_i^4}{4x_i^4} \right) - 9\alpha^3 \sum_{x_i > a} \frac{\sigma_i^4}{4x_i^4} - \alpha^4 \sum_{x_i > a} \frac{\sigma_i^4}{2x_i^4} &= 0 \end{aligned}$$

$$\approx \alpha \log(a) - (\alpha + 1) \log(x_i) + \log(\alpha) + \frac{(\alpha + 1)(\alpha + 2)\sigma_i^2}{2x_i^2}$$

Setting the derivative with respect to α to zero gives

$$\begin{aligned} \sum_{x_i > a} \left(\log(a) - \log(x_i) + \frac{1}{\alpha} + \frac{(2\alpha + 3)\sigma_i^2}{2x_i^2} \right) &= 0 \\ \frac{n_a}{\alpha} + \sum_{x_i > a} (\log(a) - \log(x_i)) + (2\alpha + 3) \sum_{x_i > a} \frac{\sigma_i^2}{2x_i^2} &= 0 \\ n_a + \alpha \sum_{x_i > a} \left(\log(a) - \log(x_i) + 3 \frac{\sigma_i^2}{2x_i^2} \right) + \alpha^2 \sum_{x_i > a} \frac{\sigma_i^2}{x_i^2} &= 0 \end{aligned}$$

Which has solution

$$\alpha = \frac{\sum_{x_i > a} \left(\log(x_i) - \log(a) - 3 \frac{\sigma_i^2}{2x_i^2} \right) \pm \sqrt{\left(\sum_{x_i > a} \left(\log(x_i) - \log(a) - 3 \frac{\sigma_i^2}{2x_i^2} \right) \right)^2 - 4n_a \sum_{x_i > a} \frac{\sigma_i^2}{x_i^2}}}{2 \sum_{x_i > a} \frac{\sigma_i^2}{x_i^2}}$$

Assuming that $\frac{\sigma_i^2}{x_i^2}$ is small. we have the approximation

$$\begin{aligned} &\sqrt{\left(\sum_{x_i > a} \left(\log(x_i) - \log(a) - 3 \frac{\sigma_i^2}{2x_i^2} \right) \right)^2 - 4n_a \sum_{x_i > a} \frac{\sigma_i^2}{x_i^2}} \\ &\approx \sum_{x_i > a} \left(\log(x_i) - \log(a) - 3 \frac{\sigma_i^2}{2x_i^2} \right) - \frac{2n_a \sum_{x_i > a} \frac{\sigma_i^2}{x_i^2}}{\sum_{x_i > a} \left(\log(x_i) - \log(a) - 3 \frac{\sigma_i^2}{2x_i^2} \right)} - \frac{2n_a^2 \left(\sum_{x_i > a} \frac{\sigma_i^2}{x_i^2} \right)^2}{\left(\sum_{x_i > a} \left(\log(x_i) - \log(a) - 3 \frac{\sigma_i^2}{2x_i^2} \right) \right)^3} \end{aligned}$$

which gives the MLE

$$\begin{aligned} \alpha &= \frac{n_a \sum_{x_i > a} \frac{\sigma_i^2}{x_i^2}}{\left(\sum_{x_i > a} \frac{\sigma_i^2}{x_i^2} \right) \left(\sum_{x_i > a} \left(\log(x_i) - \log(a) - 3 \frac{\sigma_i^2}{2x_i^2} \right) \right)} + \frac{n_a^2 \left(\sum_{x_i > a} \frac{\sigma_i^2}{x_i^2} \right)^2}{2 \left(\sum_{x_i > a} \frac{\sigma_i^2}{x_i^2} \right) \left(\sum_{x_i > a} \left(\log(x_i) - \log(a) - 3 \frac{\sigma_i^2}{2x_i^2} \right) \right)^3} \\ &= \frac{n_a}{\sum_{x_i > a} \left(\log(x_i) - \log(a) - 3 \frac{\sigma_i^2}{2x_i^2} \right)} + \frac{n_a^2 \left(\sum_{x_i > a} \frac{\sigma_i^2}{x_i^2} \right)}{\left(\sum_{x_i > a} \left(\log(x_i) - \log(a) - 3 \frac{\sigma_i^2}{2x_i^2} \right) \right)^3} \\ &\approx \frac{n_a}{\sum_{x_i > a} (\log(x_i) - \log(a))} + \left(\sum_{x_i > a} \frac{\sigma_i^2}{2x_i^2} \right) \left(\frac{3n_a}{\left(\sum_{x_i > a} (\log(x_i) - \log(a)) \right)^2} + \frac{n_a^2}{\left(\sum_{x_i > a} (\log(x_i) - \log(a)) \right)^3} \right) \end{aligned}$$

For our specific case, the normal true prior has $\tau = 1$ and $a = \Phi^{-1}(0.9) = 1.281552$. Empirically, for these parameters, $\mathbb{E}(\log(X)) = 0.538$, so $\mathbb{E}(\log(X)) - \log(a) = 0.29$, and $\mathbb{E}\left(\frac{1}{X^2}\right) = 0.37$. Therefore, the expected value of $\hat{\alpha}$ is

$$\mathbb{E}(\hat{\alpha}) = \frac{1}{0.29} + 0.3694015\mathbb{E}\sigma_i^2\left(\frac{3}{0.29^2} + \frac{1}{0.39^3}\right) = 3.45 + 28.32\mathbb{E}\sigma_i^2$$

Since σ_i follows an exponential distribution with $\lambda = 50$, so $\mathbb{E}(\sigma_i^2) = \frac{2}{50^2} = 0.0008$, which means that

$$\mathbb{E}(\hat{\alpha}) = 3.45 + 0.02 = 3.47$$

For the exponential true prior, we have

Proposition C.1. *If X follows an exponential distribution with rate λ , then the function $f(\lambda) = \mathbb{E}(\log(1 + X))$ satisfies the differential equation*

$$f'(\lambda) = f(\lambda) - \frac{1}{\lambda}$$

Proof. We have $f(\lambda) = \int_0^\infty \lambda e^{-\lambda x} \log(1 + x) dx$. This gives

$$\begin{aligned} f'(\lambda) &= \int_0^\infty e^{-\lambda x} \log(1 + x) dx - \int_0^\infty \lambda x e^{-\lambda x} \log(1 + x) dx \\ &= \int_0^\infty e^{-\lambda x} \log(1 + x) dx - \left[-e^{-\lambda x} x \log(1 + x)\right]_0^\infty - \int_0^\infty \left(\log(1 + x) + \frac{x}{1 + x}\right) e^{-\lambda x} dx \\ &= - \int_0^\infty \frac{x}{1 + x} e^{-\lambda x} dx \\ &= - \int_0^\infty \left(1 - \frac{1}{1 + x}\right) e^{-\lambda x} dx \\ &= \int_0^\infty \frac{e^{-\lambda x}}{1 + x} dx - \frac{1}{\lambda} \end{aligned}$$

On the other hand, integration by parts gives

$$\begin{aligned} f(\lambda) &= \int_0^\infty \lambda e^{-\lambda x} \log(1 + x) dx \\ &= \left[-e^{-\lambda x} \log(1 + x)\right]_0^\infty + \int_0^\infty \frac{1}{1 + x} e^{-\lambda x} dx \\ &= \int_0^\infty \frac{1}{1 + x} e^{-\lambda x} dx \end{aligned}$$

Substituting this into the previous equation gives

$$f'(\lambda) = f(\lambda) - \frac{1}{\lambda}$$

□

Proposition C.2. *If X follows an exponential distribution with rate λ , then the function $g(\lambda) = \mathbb{E}\left(\frac{1}{(1+X)^2}\right)$ satisfies the differential equation*

$$g'(\lambda) = \left(1 + \frac{2}{\lambda}\right)g(\lambda) - 1$$

Proof. We have $g(\lambda) = \int_0^\infty \frac{\lambda e^{-\lambda x}}{(1+x)^2} dx$. This gives

$$\begin{aligned} g'(\lambda) &= \int_0^\infty e^{-\lambda x} \frac{1 - \lambda x}{(1+x)^2} dx \\ &= (1 + \lambda) \int_0^\infty \frac{e^{-\lambda x}}{(1+x)^2} dx - \lambda \int_0^\infty \frac{e^{-\lambda x}}{(1+x)} dx \end{aligned}$$

On the other hand, integration by parts gives

$$\begin{aligned} \int_0^\infty \frac{\lambda e^{-\lambda x}}{1+x} dx &= \left[-\frac{e^{-\lambda x}}{1+x} \right]_0^\infty - \int_0^\infty \frac{e^{-\lambda x}}{(1+x)^2} dx \\ &= 1 - \frac{g(\lambda)}{\lambda} \end{aligned}$$

This gives us

$$\begin{aligned} g'(\lambda) &= \frac{(1 + \lambda)}{\lambda} g(\lambda) - \left(1 - \frac{g(\lambda)}{\lambda}\right) \\ &= \left(1 + \frac{2}{\lambda}\right)g(\lambda) - 1 \end{aligned}$$

□

This means that for an exponential with parameter $\lambda = 1$ and cut-off $a = \log(10)$, $Z = \frac{X}{a} - 1$ follows an exponential distribution with rate a , so $\log(X) - \log(a) = \log(1 + Z)$, so its expected value is $f(a)$, where f is the solution to

$$f'(\lambda) = f(\lambda) - \frac{1}{\lambda}$$

Similarly, $X_i^{-2} = (a(1 + Z))^{-2}$, so $\mathbb{E}(X_i^{-2}) = a^{-2}g(a)$. The expected value of $\hat{\alpha}$ is then

$$\frac{1}{f(\log(10))} + \frac{\mathbb{E}(\sigma_i^2)g(\log(10))}{2\log(10)^2} \left(\frac{3}{f(\log(10))^2} + \frac{1}{f(\log(10))^3} \right)$$

Numerically, we find $f(\log(10)) = 0.3239$ and $g(\log(10)) = 0.5853$. Substituting these values into the equation gives

$$\hat{\alpha} = \frac{1}{0.3239} + \mathbb{E}(\sigma_i^2) \frac{0.5853}{2\log(10)^2} \left(\frac{3}{0.3239^2} + \frac{1}{0.3239^3} \right) = 3.087 + 3.203\mathbb{E}(\sigma_i^2) \approx 3.151$$

References

- [1] Aitkin, M. and Longford, N. (1986) Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)* 149, 1–43
- [2] Robert E Bechhofer (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, 16–39,
- [3] Gelman, A. and Price, P. N. (1999). All maps of parameter estimates are misleading. *Statistics in Medicine* 18, 3221–3234
- [4] Gupta, S. S. (1956) On a decision rule for a problem in ranking means. *PhD thesis, University of North Carolina at Chapel Hill*
- [5] Gupta, S. S. and Hsiao, P. (1983) Empirical Bayes rules for selecting good populations. *Journal of Statistical Planning and Inference* 8, 87–101
- [6] Henderson, N. C. and Newton, M. A. (2015), Making the cut: improved ranking and selection for large-scale inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. doi: 10.1111/rssb.12131
- [7] Laird, N. (1978) Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Journal of the American Statistical Association* 73, 805–811
- [8] Laird, N. M. and Louis, T. A. (1989) Empirical Bayes ranking methods. *Journal of Educational and Behavioral Statistics* 14, 29–46,
- [9] Lin, R., Louis, T. A., Paddock, S. M. and Ridgeway, G. (2006) Loss function based ranking in two-stage, hierarchical models. *Bayesian Analysis (Online)*, 1(4):915,
- [10] Morris, A. P., B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segre, V. Steinthorsdottir, R. J. Strawbridge, H. Khan, H. Grallert, A. Mahajan, et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* 44, 981–990.
- [11] West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H, Olson, J. A. Jr., Marks, J. R. and Nevins, J. R. (2001), Predicting the clinical status of human breast cancer by using gene expression profiles, *Proceedings of the National Academy of Sciences* (98), 11462–11467